HAL
open science

Estimation robuste de courbes de consommmation électrique moyennes par sondage pour de petits domaines en présence de valeurs manquantes

Anne de Moliner

▶ **To cite this version:**

Anne de Moliner. Estimation robuste de courbes de consommmation électrique moyennes par sondage pour de petits domaines en présence de valeurs manquantes. Mathématiques générales [math.GM]. Université Bourgogne Franche-Comté, 2017. Français. NNT : 2017UBFCK021 . tel-01820617

HAL Id: tel-01820617

https://theses.hal.science/tel-01820617

Submitted on 22 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE PREPAREE A L'INSTITUT DE MATHEMATIQUES DE BOURGOGNE

Ecole doctorale 553 Carnot Pasteur

Doctorat de Mathématiques

Par

Anne DE MOLINER

Estimation robuste de courbes de consommation électrique moyennes par sondage pour de petits domaines en présence de valeurs manquantes

Jury

Anne Ruiz-Gazen	Université Toulouse Capitole	Présidente du jury
David Haziza	Université de Montréal	Rapporteur
Jean-Michel Poggi	Université Paris Descartes	Rapporteur
Hervé Cardot	Université Bourgogne	Directeur de thèse
Camelia Goga	Université de Franche-Comté	Co-directrice de thèse
Jérôme Cubillé	EDF R&D	Invité

THESE

Table des matières

Table des matières 1

Liste des tableaux 7

1 Introduction 7

2 État de l'art sur l'estimation par sondage pour des données fonctionnelles 13
- 2.1 Données fonctionnelles . . . 13
  - 2.1.1 Cadre de travail . . . 14
    - Hypothèses et notations sur les données fonctionnelles . . . 14
    - Application au contexte EDF . . . 15
  - 2.1.2 Projection de courbes en dimension finie . . . 17
  - 2.1.3 Réduction de dimension : Analyse en Composantes Principales fonctionnelle . . . 18
  - 2.1.4 Régression linéaire pour des données fonctionnelles . . . 20
- 2.2 Sondages . . . 21
  - 2.2.1 Notations sur les sondages . . . 21
  - 2.2.2 Estimateur de Horvitz-Thompson . . . 23
  - 2.2.3 Estimateur de Horvitz-Thompson pour des données fonctionnelles 25
  - 2.2.4 Prise en compte de variables explicatives . . . 26
    - Au niveau de l'échantillonnage : sondage stratifié . . . 26
    - Au niveau de l'estimation : estimateur par calage fonctionnel . . . 28
    - Cas particulier du calage : estimateur de Hájek . . . 31
    - Approche basée sur le modèle . . . 32
  - 2.2.5 Estimation de la variance sous le plan . . . 33
    - Linéarisation en sondages . . . 34
    - Bootstrap en sondages . . . 35

3 Estimation par sondage de courbes moyennes ou totales de consommation électrique robuste aux unités influentes 37
- 3.1 Contexte et introduction . . . 37
- 3.2 Robustesse en statistique . . . 40
- 3.3 Biais conditionnel pour des courbes . . . 42
- 3.4 Estimation robuste de la courbe totale . . . 43
  - 3.4.1 Estimation robuste de courbe totale, instant par instant . . . 44
    - Démarche de construction d'estimateurs robustes de courbes totales, instant par instant . . . 44

Illustration sur des données réelles 46
Avantages et inconvénients de la méthode 47
3.4.2 Estimation robuste sur une base de projection 48
Analyse en Composantes Principales Sphériques 49
Des scores moyens à la courbe moyenne 51
Estimation sur la base de l'ACP sphérique : illustration sur des données réelles 53
3.4.3 Troncature fonctionnelle basée sur la notion de profondeur 54
Notion de profondeur pour des données fonctionnelles 54
Modified Band Depth 54
Définition de profondeurs basées sur la projection en dimension finie. 55
Troncature fonctionnelle à partir de la notion de profondeur 56
Illustration sur des données réelles 58
3.4.4 Proposition de nouveau critère pour le choix des constantes d'ajustement 59
Nouveau critère pour l'estimation robuste instant par instant 60
Nouveau critère pour l'estimation robuste sur une base de projection 61
Nouveau critère pour l'estimation robuste par troncature fonctionnelle 61
3.5 Estimation d'erreur quadratique moyenne (EQM) pour les estimateurs robustes de courbes 62
3.5.1 Estimateurs explicites d'erreur quadratique moyenne 62
Estimateurs explicites d'EQM pour l'estimation robuste instant par instant 63
Estimateurs explicites d'EQM pour la troncature dans l'espace des composantes principales 63
Estimateurs explicites d'EQM pour la troncature fonctionnelle . . . 65
3.5.2 Estimateurs d'EQM par bootstrap populationnel 65
3.5.3 Estimateurs d'EQM par bootstrap généralisé 66
Construction des estimateurs non robustes dans les réplications . . 68
Estimation des biais conditionnels dans les réplications 68
Construction des estimateurs robustes dans les réplications 69
3.6 Application sur des jeux de données réelles 70
3.6.1 Présentation du jeu de données 70
3.6.2 Tests des estimateurs robustes de courbes moyennes 72
Protocole de test 72
Indicateurs de qualité 74
Résultats pour les estimateurs de courbe moyenne 75
3.6.3 Tests des estimateurs d'erreur quadratique moyenne 78
Procédure de test 80
Indicateurs de qualité 81
Résultats pour l'estimation d'erreur quadratique moyenne 81
3.7 Conclusions sur la robustesse 95
3.7.1 Conclusions méthodologiques 95

3.7.2 Cas d'application des méthodes robustes 96
3.7.3 Perspectives 96

4 Estimation de courbes de consommation électrique moyennes ou totales par sondage pour de petits domaines 99
4.1 Contexte et introduction 99
4.1.1 Notations et cadre de travail 102
4.2 Estimation de courbes moyennes pour des petits domaines, dans l'approche basée sur le plan de sondage 103
4.3 Estimation de courbes moyennes pour des petits domaines dans l'approche basée sur un modèle 104
4.3.1 Modèles linéaires mixtes au niveau unité pour des données fonctionnelles 105
4.3.2 Régression linéaire fonctionnelle 108
Estimateurs de courbes moyennes de domaines par régression linéaire fonctionnelle 108
Estimation rapide des coefficients du modèle à l'aide de l'algorithme du calage 109
4.3.3 Agrégation de prédictions par arbres de régression pour des courbes 110
Approche prédictive pour des domaines et estimation non paramétrique 110
Arbres de régression pour des données fonctionnelles 111
4.3.4 Agrégation de prédictions par forêts aléatoires pour des courbes . . 113
4.4 Estimation de courbes moyennes robuste aux unités influentes pour des petits domaines 115
4.4.1 Approches basées sur le biais conditionnel 115
Approches basées sur le biais conditionnel, pour les méthodes basées sur le plan 116
Biais conditionnels dans l'approche basée sur un modèle 117
Estimation robuste par modèles linéaires mixtes sur une base de projection 118
Estimation robuste par régression linéaire fonctionnelle 120
Estimation robuste par approche prédictive et arbres de régression pour des courbes 121
Estimation robuste par approche prédictive et forêts aléatoires pour des courbes 122
4.4.2 Estimateurs REBLUP sur composantes principales sphériques . . . 123
4.5 Application à des courbes de consommation électrique 125
4.5.1 Présentation du jeu de données 125
4.5.2 Protocole de test 126
4.5.3 Résultats et conclusion des tests 129
4.6 Conclusions sur les petits domaines 136
4.6.1 Conclusions méthodologiques 136
4.6.2 Perspectives 136

5 Estimation de courbes moyennes par sondage en présence de valeurs manquantes 139
5.1 Introduction et contexte 139
5.1.1 Notations 141
5.2 Estimation de courbe moyenne ou totale par repondération par lissage à noyau 142
5.2.1 Estimateur de courbe moyenne par repondération par lissage à noyau 142
Lissage à noyau de la courbe moyenne sur l'ensemble de la population 143
Estimateurs non paramétriques en sondages 143
Estimateurs non paramétriques en sondage en présence de non réponse 144
Estimation des probabilités de réponse $\vartheta_c(t_l)$ et $\vartheta_c(t_l, t_{l'})$. 145
Cas particulier du sondage stratifié avec groupes de réponses homogènes par strate 145
Choix de la fenêtre 147
5.2.2 Estimation de variance pour l'estimateur à noyau par repondération 147
L'erreur d'approximation et le biais sont négligeables 148
Approximation de variance pour l'estimateur de Horvitz-Thompson 148
Approximation de variance pour les estimateurs de Hájek 149
Formule de variance pour le sondage stratifié 151
Comparaison des variances pour le sondage stratifié 152
5.3 Estimation de courbes moyennes ou totales en présence de valeurs manquantes par imputation 153
5.3.1 Imputation par des estimateurs à noyau de la courbe de charge . . 153
5.3.2 Imputation par les plus proches voisins 154
5.3.3 Interpolation linéaire de la différence à la moyenne 155
5.3.4 Approche unifiée pour l'approximation de variance des estimateurs par imputation 157
Imputation linéaire 158
Modèle de superpopulation et décomposition générale de la variance. 159
Estimation de variance pour l'imputation par les plus proches voisins 161
Estimation de variance pour l'imputation basée sur l'interpolation linéaire 163
Estimateur de variance pour l'imputation par l'estimateur à noyau 164
5.4 Application aux données réelles de consommation électriques 164
5.4.1 Données, plans de sondage et scénarios non réponse 164
5.4.2 Estimateurs comparés et indicateurs de performance 166
5.4.3 Aspects pratiques de l'implémentation des estimateurs de courbe moyenne ou totale ainsi que des estimateurs de variance associés . 167
5.4.4 Résultats 168
5.5 Conclusions sur l'estimation de courbes en présence de valeurs manquantes 180

5.5.1 Conclusions méthodologiques 180
5.5.2 Perspectives 181
Liste des tableaux

Liste des tableaux

3.1 MSE de l'estimateur non robuste par taille d'échantillon (taille) pour les différentes stratégies d'estimation 77
3.2 Indicateurs de qualité pour le sondage aléatoire simple (SAS HT) et l'estimateur de Horvitz-Thompson en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode) 86
3.3 Indicateurs de qualité pour le sondage aléatoire simple avec calage (SAS calage) en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode) 87
3.4 Indicateurs de qualité pour le sondage stratifié sans strata jumper (STR HT) en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode) 88
3.5 Indicateurs de qualité pour le sondage stratifié avec 10% de strata jumpers en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode) 89
3.6 Indicateurs de qualité pour le sondage stratifié avec 20% de strata jumpers en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode) 90
3.7 Biais relatifs et temps de calcul des différentes méthodes d'estimation (sondage aléatoire simple, 100 unités et estimateur de Horvitz-Thompson) 91
3.8 Biais relatifs et temps de calcul des différentes méthodes d'estimation (sondage stratifié, 100 unités et estimateur de Horvitz-Thompson) 91
3.9 Biais relatifs et temps de calcul des différentes méthodes d'estimation (sondage aléatoire simple, 100 unités, et estimateur par calage) 92

4.1 Différentes méthodes d'estimation testées. 127
4.2 Paramétrage des arbres et forêts aléatoires. 127
4.3 Moyennes des indicateurs de performances des méthodes (RB, MSE, RE), pour l'ensemble des instants de discrétisation et des domaines, en séparant le domaine non échantillonné des autres. 134
4.4 Temps de calcul moyen en secondes des différents estimateurs pour un échantillon de 200 courbes de 177 points 135

5.1 RB (%) pour chaque méthode d'estimation et chaque scénario (partie 1). . 170
5.2 RB (%) pour chaque méthode d'estimation et chaque scénario (partie 2). . 170
5.3 MSE pour chaque méthode d'estimation et chaque scénario (partie 1). . . 171
5.4 MSE pour chaque méthode d'estimation et chaque scénario (partie 2). . . 171
5.5 RE (%) pour chaque méthode d'estimation et chaque scénario (partie 1). . 171

5.6 RE (%) pour chaque méthode d'estimation et chaque scénario (partie 2). . 172
5.7 Biais relatifs (en %) des estimateurs de variance, pour le sondage stratifié et le scénario mixte. 172
5.8 Temps de calcul (en secondes) pour les différentes méthodes d'estimation, avec et sans calcul de variance. 172

Remerciements

Tout d'abord, je tiens à remercier mes directeurs de thèse Hervé Cardot et Camelia Goga d'avoir accepté d'encadrer cette thèse, et surtout pour leurs précieux conseils pendant ces quatre années ainsi que leurs relectures nombreuses et patientes de ce manuscrit et des articles. Je remercie également mon référent industriel, Jérôme Cubillé pour son aide.

Je tiens à remercier les rapporteurs, Jean-Michel Poggi et David Haziza pour le temps consacré à la lecture de ce long manuscrit ainsi que leurs remarques pertinentes. Merci également à Anne Ruiz-Gazen, d'avoir accepté de faire partie de mon jury de thèse.

Je voudrais également remercier mes cheffes, Christine, Caroline et Maud, qui m'ont permis de me lancer dans ce projet et m'ont dégagé du temps pour pouvoir m'y consacrer.

Merci aux équipes SOAD et E74, en particulier à ceux qui ont contribué de plus près à la thèse : Benoît pour la maquette de Valenciennes, Christophe pour le Courbotree, Cécile, Isabelle et Sophie, les cheffes de projet qui ont suivi ce travail et ses applications et m'ont orientée sur les aspects industriels. Merci à France pour son aide logistique dans la préparation de la soutenance. Je remercie aussi mes autres collègues et anciens collègues pour l'ambiance de travail très agréable, qui va beaucoup me manquer, pour les diamants, les cafés du matin, et les discussions (parfois) scientifiques, notamment Bérénice, Lou, Guillaume, Bruno, Manel, Amandine, Aurélien, Laurent, Marie-Anne et Jiali.

Merci à Vincent et à Cyril pour le temps passé en conférence. Enfin merci à Alain et Jérôme pour leurs conseils lorsque j'ai voulu me lancer dans cette aventure.

Enfin je remercie mes proches, ceux qui m'ont accompagné et m'ont aidé à me changer les idées tout au long de ces quatre ans. Je remercie mes amis, en particulier Francois, Adé, Thu-Van, Greg, Meriem, Pierre, Samira, Xavier, Pierre-Hugues, Mathieu, Antoine, et Mickael. Merci à Astrid pour son soutien lointain mais important. Merci à l'équipe du SBIM : Loic, Sammy et Alexandra. Merci à Julien, Geoffrey, Etienne, Elina et Aurélien, pour les américains fricadelles sauce pitta. Merci à ceux que j'ai oubliés et qui ne m'en voudront pas.

Enfin, je tiens à exprimer aussi toute ma gratitude envers ma famille, mes parents et mes beaux parents, qui m'ont soutenu pendant ces quatre longues années. En particulier, merci à Odile pour son aide culinaire. Merci à Gautier pour sa patience. Enfin, je remercie Gaspard, sans qui cette thèse aurait été finie beaucoup plus rapidement.

Résumé

Dans cette thèse, nous nous intéressons à l'estimation robuste de courbes moyennes ou totales de consommation électrique par sondage en population finie, pour l'ensemble de la population ainsi que pour des petites sous-populations, en présence ou non de courbes partiellement inobservées.

En effet, de nombreuses études réalisées dans le groupe EDF, que ce soit dans une optique commerciale ou de gestion du réseau de distribution par Enedis, se basent sur l'analyse de courbes de consommation électrique moyennes ou totales, pour différents groupes de clients partageant des caractéristiques communes. L'ensemble des consommations électriques de chacun des 35 millions de clients résidentiels et professionnels français ne pouvant être mesurées pour des raisons de coût et de protection de la vie privée, ces courbes de consommation moyennes sont estimées par sondage à partir de panels. Nous prolongeons les travaux de Lardin (2012) sur l'estimation de courbes moyennes par sondage en nous intéressant à des aspects spécifiques de cette problématique, à savoir l'estimation robuste aux unités influentes, l'estimation sur des petits domaines, et l'estimation en présence de courbes partiellement ou totalement inobservées.

Pour proposer des estimateurs robustes de courbes moyennes, nous adaptons au cadre fonctionnel l'approche unifiée d'estimation robuste en sondages basée sur le biais conditionnel proposée par Beaumont et al. (2013). Pour cela, nous proposons et comparons sur des jeux de données réelles trois approches : l'application des méthodes usuelles sur les courbes discrétisées, la projection sur des bases de dimension finie (Ondelettes ou Composantes Principales de l'Analyse en Composantes Principales Sphériques Fonctionnelle en particulier) et la troncature fonctionnelle des biais conditionnels basée sur la notion de profondeur d'une courbe dans un jeu de données fonctionnelles. Des estimateurs d'erreur quadratique moyenne instantanée, explicites et par bootstrap, sont également proposés.

Nous traitons ensuite la problématique de l'estimation sur de petites sous-populations. Dans ce cadre, nous proposons trois méthodes : les modèles linéaires mixtes au niveau unité appliqués sur les scores de l'Analyse en Composantes Principales ou les coefficients d'ondelettes, la régression fonctionnelle et enfin l'agrégation de prédictions de courbes individuelles réalisées à l'aide d'arbres de régression ou de forêts aléatoires pour une variable cible fonctionnelle. Des versions robustes de ces différents estimateurs sont ensuite proposées en déclinant la démarche d'estimation robuste basée sur les biais conditionnels proposée précédemment.

Enfin, nous proposons quatre estimateurs de courbes moyennes en présence de courbes partiellement ou totalement inobservées. Le premier est un estimateur par repondération par lissage temporel non paramétrique adapté au contexte des sondages

et de la non réponse et les suivants reposent sur des méthodes d'imputation. Les portions manquantes des courbes sont alors déterminées soit en utilisant l'estimateur par lissage précédemment cité, soit par imputation par les plus proches voisins adaptée au cadre fonctionnel ou enfin par une variante de l'interpolation linéaire permettant de prendre en compte le comportement moyen de l'ensemble des unités de l'échantillon. Des approximations de variance sont proposées dans chaque cas et l'ensemble des méthodes sont comparées sur des jeux de données réelles, pour des scénarios variés de valeurs manquantes.

Mots Clés Arbres de régression, biais conditionnels, données fonctionnelles, données manquantes, estimation sur petits domaines, estimateurs à noyau, forêts aléatoires, modèles linéaires mixtes, plus proches voisins, robustesse, sondage ...

Abstract

In this thesis, we address the problem of robust estimation of mean or total electricity consumption curves by sampling in a finite population for the entire population and for small areas. We are also interested in estimating mean curves by sampling in presence of partially missing trajectories.

Indeed, many studies carried out in the French electricity company EDF, for marketing or power grid management purposes, are based on the analysis of mean or total electricity consumption curves at a fine time scale, for different groups of clients sharing some common characteristics. Because of privacy issues and financial costs, it is not possible to measure the electricity consumption curve of each customer so these mean curves are estimated using samples. In this thesis, we extend the work of Lardin (2012) on mean curve estimation by sampling by focusing on specific aspects of this problem such as robustness to influential units, small area estimation and estimation in presence of partially or totally unobserved curves.

In order to build robust estimators of mean curves we adapt the unified approach to robust estimation in finite population proposed by Beaumont et al. (2013) to the context of functional data. To that purpose we propose three approaches : application of the usual method for real variables on discretised curves, projection on Functional Spherical Principal Components or on a Wavelets basis and thirdly functional truncation of conditional biases based on the notion of depth. These methods are tested and compared to each other on real datasets and Mean Squared Error estimators are also proposed.

Secondly we address the problem of small area estimation for functional means or totals. We introduce three methods : unit level linear mixed model applied on the scores of functional principal components analysis or on wavelets coefficients, functional regression and aggregation of individual curves predictions by functional regression trees or functional random forests. Robust versions of these estimators are then proposed by following the approach to robust estimation based on conditional biais presented before.

Finally, we suggest four estimators of mean curves by sampling in presence of partially or totally unobserved trajectories. The first estimator is a reweighting estimator where the weights are determined using a temporal non parametric kernel smoothing adapted to the context of finite population and missing data and the other ones rely on imputation of missing data. Missing parts of the curves are determined either by using the smoothing estimator presented before, or by nearest neighbours imputation adapted to functional data or by a variant of linear interpolation which takes into account the mean trajectory of the entire sample. Variance approximations are proposed for each method and all the estimators are compared to each other on real datasets for

various missing data scenarios.

Key Words conditional bias, functional data, kernel estimators, missing data, linear mixed models, nearest neighbours, random forests, regression trees, robustness, small area estimation, survey sampling . . .

Chapitre 1

Introduction

Dans cette thèse, on s'intéresse à l'estimation robuste de courbes moyennes ou totales de consommation électrique par sondage en population finie, pour l'ensemble de la population ainsi que pour des petites sous-populations, en présence ou non de courbes partiellement observées.

De nombreuses études réalisées à EDF R&D se basent sur l'analyse de courbes de consommation électrique moyennes ou totales, pour différents groupes de clients partageant des caractéristiques communes (par exemple des équipements électriques similaires ou un tarif commun). Ces estimations peuvent trouver de très nombreuses applications métier, que ce soit dans une optique de marketing et de connaissance client pour la direction commerciale de EDF ou dans le cadre de la gestion du réseau électrique par Enedis.

Par exemple, dans une optique de connaissance client, ces analyses peuvent permettre de quantifier l'impact des équipements ou tarifs sur la consommation électrique, en comparant les courbes de consommation électrique de différents groupes de clients possédant différentes caractéristiques. En outre, dans une optique de prospective, l'étude de ces courbes moyennes peut permettre d'établir des scénarios d'évolution des consommations globales en fonction des évolutions des usages.

Par ailleurs, dans le domaine de la gestion du réseau de distribution d'électricité, Enedis se doit d'assurer à tout moment l'équilibre entre offre et demande d'électricité ; pour cela, il est nécessaire de savoir quelle quantité d'énergie a été consommée sur le réseau à chaque instant et, comme les consommations ne sont pas mesurées à un pas de temps fin sur l'ensemble du réseau pour l'ensemble des clients, il est nécessaire d'estimer les courbes de consommation à la maille de chaque fournisseur d'électricité.

Afin d'estimer les courbes de consommation électrique moyennes de nos différentes populations d'intérêt, nous disposons de panels de plusieurs milliers voire dizaines de milliers de clients, sélectionnés selon un plan de sondage aléatoire, et dont on mesure la courbe de consommation électrique individuelle globale (tous usages confondus) à un pas de temps demi-horaire pendant de longues périodes, souvent des années. On leur adresse éventuellement aussi un questionnaire concernant leurs caractéristiques socio démographiques, mais aussi celles de leurs logements et de leurs équipements et usages électriques. Enfin, on dispose en outre de variables explicatives associées provenant de la facturation (tarif, option et puissance souscrite, consommations annuelles).

Le déploiement de compteurs communicants Linky et PME-PMI actuellement en cours chez l'ensemble des clients résidentiels et professionnels français facilitera grandement la constitution des panels et la collecte de courbes de consommation électrique. Néanmoins, même à l'issue de ce déploiement, nous continuerons à réaliser nos estimations de courbes moyennes par sondage plutôt que de traiter l'intégralité des données. En effet, la remontée automatique de l'ensemble des courbes de consommation collectées par ces compteurs au pas dix ou trente minutes représenterait un coût très important. De plus, l'utilisation par un fournisseur ou distributeur d'électricité de ces données est soumise à l'accord explicite du client du fait de la nature personnelle de cette information, ce qui empêche la collecte systématique et induit des coûts supplémentaires. Enfin, dans un contexte d'études de marketing et de connaissance de la clientèle, il est nécessaire de collecter des informations relatives aux caractéristiques socio-démographiques ainsi qu'aux équipements électriques et aux caractéristiques du bâti pour les clients du panel afin d'analyser leur impact sur les courbes de consommation électrique associées. Cela rend donc nécessaire la passation de questionnaires et par conséquent la collecte par sondage.

La question de l'estimation de courbes moyennes ou totales par sondage en population finie, en particulier pour des consommations électriques, a déjà été traitée dans la littérature. En effet, Cardot et al. (2010) a proposé un estimateur de Horvitz-Thompson de la courbe moyenne. Dans Cardot and Josserand (2011), les propriétés de convergence uniforme de cet estimateur ont été démontrées et des méthodes d'estimation de bandes de confiance ont été proposés. Ces travaux ont ensuite été poursuivis avec Cardot et al. (2013b) qui propose un estimateur assisté par un modèle pour l'estimation de la courbe moyenne par sondage en population finie en adaptant l'estimateur de Särndal et al. (2003) au cadre des données fonctionnelles. Dans ce même article, différentes stratégies d'estimation de la courbe moyenne en présence d'information auxiliaire sont comparées sur un jeu de données de courbes de consommation électrique et des méthodes de construction des bandes de confiance associées sont proposées.

Cette thèse s'intéresse donc tout particulièrement à des aspects spécifiques de l'estimation par sondage en population finie qui n'ont pas encore été abordés dans le cadre des données fonctionnelles : l'estimation robuste aux unités influentes, l'estimation pour de petites sous-populations aussi appelées petits domaines et enfin l'estimation en présence de courbes partiellement observées. Les deux premières problématiques seront traitées séparément puis conjointement. Dans le contexte des sondages, hors du cadre des données fonctionnelles, ces différents points ont déjà été étudiés. En effet, Beaumont et al. (2013) ont proposé une approche unifiée pour l'estimation robuste en sondages en se basant sur une mesure d'influence appelée biais conditionnel, introduite par Muñoz-Pichardo et al. (1995) puis reprise par Moreno-Rebollo et al. (1999) dans le cadre de l'estimation par sondage en population finie. Par ailleurs, l'estimation sur de petites sous-populations a fait l'objet d'une large littérature, très bien exposée dans l'ouvrage Rao and Molina (2015). Enfin, la question de la non réponse partielle en sondages, qui est très similaire à notre problématique des courbes partiellement inobservées a été abondamment étudiée et on pourra se référer notamment à Särndal and Lundström (2005) ou Haziza (2009) ainsi qu'aux références citées dans ces articles. Pour l'ensemble des sujets traités, notre objectif sera donc d'adapter ou

d'étendre les méthodes existantes à notre contexte particulier dans lequel notre variable d'intérêt est une courbe, de façon à tirer le meilleur parti possible des fortes corrélations temporelles de notre problématique pour améliorer la précision de nos estimateurs.

Ces travaux sont également basés sur des méthodes issues d'autres champs de la statistique que les sondages. En effet, pour exploiter les dépendances temporelles de nos données, on utilisera des méthodes et concepts issus de l'analyse des données fonctionnelles. Cette branche de la statistique, dont l'objet est l'étude de données de type courbes ou fonctions, a connu un essor important au cours des années 1990 et 2000 avec le développement des capacités de traitement et de stockage des données mais surtout la multiplication des données issues de capteurs dans différents domaines scientifiques et techniques tels que la médecine ou l'industrie notamment. Les ouvrages de Ramsay and Silverman (2005), Ferraty and Vieu (2006) ainsi que l'article de Cuevas (2014) constituent des références sur ce sujet. Dans le cadre de cette thèse, nous utiliserons tout particulièrement des outils issus de Ramsay and Silverman (2005) tels que l'Analyse en Composantes Principales Fonctionnelle ou encore la Régression Fonctionnelle.

Cette thèse se situe donc au croisement de trois domaines de la statistique : l'estimation par sondage en population finie, la robustesse et l'analyse des données fonctionnelles, qui ont chacun fait l'objet d'une large littérature, ont parfois été étudiés deux à deux mais n'ont à notre connaissance jamais été considérés tous les trois ensemble.

Dans le chapitre 2 nous introduisons quelques notions et concepts issus de la littérature sur l'estimation par sondage et les données fonctionnelles. Cet état de l'art général sera complété en début de chaque chapitre par quelques références bibliographiques complémentaires propres à chaque volet de nos travaux.

Dans le chapitre 3, nous cherchons à construire des estimateurs de courbes moyennes ou totales de consommation électrique robustes à la présence d'individus atypiques. En effet, les consommations électriques ayant par nature une distribution très asymétrique, avec une grande majorité de petits clients et quelques très gros, le fait que ces derniers soient ou non sélectionnés dans l'échantillon peut avoir une influence notable sur les estimateurs. En outre, comme nous sommes dans le cadre de l'estimation par sondage en population finie, l'influence d'une unité sur un estimateur de moyenne peut également être importante du fait d'un poids de sondage fort, ou encore dans le cas d'un sondage stratifié parce qu'elle a été classée à tort dans la mauvaise strate à cause d'une information auxiliaire erronée ou mal actualisée. L'impact de ces unités influentes pourra être d'autant plus grand que les populations d'intérêt étudiées sont petites, et donc que l'on dispose de peu d'individus leur appartenant dans les panels. Il apparaît donc pertinent de développer des estimateurs robustes, c'est-à-dire peu sensibles à la présence des unités influentes. Pour cela, nous proposons et mettons en œuvre trois méthodes qui permettent d'étendre les travaux de Beaumont et al. (2013) au cadre des données fonctionnelles. La première consiste à travailler indépendamment sur les valeurs des courbes à chaque instant de discrétisation. La seconde consiste à se ramener de notre problème fonctionnel de dimension infinie à un ensemble de problèmes décorrélés d'estimation de moyenne de variables réelles que l'on

peut résoudre séparément sans briser la structure temporelle de notre problématique. Par exemple on peut utiliser une Analyse en Composantes Principales Sphériques proposée par Locantore et al. (1999), ou encore une projection sur une base d'ondelettes (voir Mallat (1999)). Enfin, une troisième approche consiste à utiliser la notion de profondeur d'une courbe, qui est une mesure d'atypicité, pour proposer une troncature purement fonctionnelle des biais conditionnels. Une comparaison sur des données réelles de ces différentes méthodes nous conduit à privilégier la seconde. Des approximations de l'erreur quadratique moyenne des différents estimateurs sont également proposés. Enfin, on suggère un nouveau critère de choix du seuil optimal de troncature qui permet le compromis entre biais et variance.

Dans le chapitre 4, nous cherchons à construire des estimateurs de courbes de consommation électrique moyennes pour des petits domaines, c'est-à-dire de petites sous-populations disjointes. En effet, l'arrivée des compteurs communicants, en diminuant les coûts d'acquisition des données de consommation électrique, va nous permettre d'accroître la taille de nos panels. Il sera alors possible de réaliser des estimations de courbes moyennes de consommation électrique non seulement à la maille de l'ensemble du territoire français, mais aussi pour des zones géographiques plus fines telles que des régions, des départements, des villes voire des quartiers. Ces estimations nous permettront de répondre à des besoins émergents tels que l'intégration sur le réseau des énergies renouvelables actuellement en fort développement ou encore de développer des services énergétiques à destination des collectivités territoriales basés sur l'analyse d'une estimation de leur consommation électrique. Dans ce contexte, du fait de la faible taille de chacun de ces domaines, les estimations basées uniquement sur les unités appartenant au domaine concerné sont très instables, voire impossibles si aucune unité du domaine n'est présente dans le panel.

L'idée directrice commune à de nombreuses méthodes de la littérature sur l'estimation sur petits domaines est de postuler l'existence d'un modèle commun à l'ensemble des sous-populations, qui va nous permettre d'exploiter conjointement l'ensemble des données pour consolider les estimations à travers l'utilisation de modèles liant la variable d'intérêt et l'ensemble des variables explicatives disponibles. Dans notre contexte, cette information auxiliaire sera celle disponible au niveau des individus à partir notamment de la facturation mais aussi au niveau des domaines d'intérêt à partir de données ouvertes. En particulier on exploitera les caractéristiques sociodémographiques des quartiers fournies par l'INSEE dans les bases de données IRIS.

Ici encore, nous proposons différentes approches pour adapter les méthodes existantes au cadre fonctionnel. La première repose sur une idée similaire à celles proposées dans le cadre de la robustesse : il s'agit de se ramener dans une base de dimension finie, par exemple par une Analyse en Composantes Principales Fonctionnelle ou une projection sur une base d'ondelettes puis d'utiliser des modèles linéaires mixtes au niveau unité, très usuels dans le contexte des petits domaines (Battese et al. (1988)), sur chacun des vecteurs de cette base. Ainsi dans le cas de l'ACP fonctionnelle, on modélise la valeur des scores des unités pour chaque composante principale puis on en déduit une estimation des scores moyens de chaque sous-population pour chaque composante principale et enfin on reconstitue une estimation de l'approximation de la courbe moyenne des différents domaines à partir de ces scores moyens. On présente également une variante de cette méthode reposant sur un modèle de ré-

gression fonctionnelle, estimée en se servant de l'algorithme du calage bien connu en sondages. Enfin, on développe également des méthodes non paramétriques de prédiction de données fonctionnelles basées sur des arbres de régression et des forêts aléatoires, en prolongeant les travaux de Stéphan and Cogordan (2009). On agrège ensuite les prédictions de courbes individuelles pour en déduire des estimateurs de courbes moyennes selon l'approche prédictive proposée par Valliant et al. (2000).

En pratique, on est souvent confrontés simultanément aux problématiques de la robustesse et de l'estimation sur des petits domaines. On propose donc également des estimateurs de courbes moyennes robustes aux unités influentes pour des petits domaines. Dans cette optique, nous avons choisi de décliner la démarche proposée au chapitre 3 pour chacun de nos estimateurs de courbes moyennes de petits domaines. On doit donc estimer les biais conditionnels des unités échantillonnées pour chacun des estimateurs, ce que l'on fait en se basant sur les travaux de Jiongo et al. (2013) en ce qui concerne les modèles linéaires mixtes et sur Bar-Hen and Poggi (2016) pour les méthodes non paramétriques. Nous avons également testé une solution alternative consistant à projeter les courbes dans un espace de dimension finie puis à utiliser l'estimateur REBLUP (Robust Empirical Best Linear Unbiased Predictor) de modèles linéaires mixtes au niveau unité proposé par Sinha and Rao (2009) pour chacun des vecteurs de base. L'ensemble des méthodes proposées dans ce chapitre ont fait l'objet de tests sur des données réelles.

Du fait de problèmes techniques pouvant survenir tout au long de la chaîne d'acquisition, de collecte et de stockage de la donnée, il peut arriver que tout ou partie de certaines courbes des panels soient manquantes. Dans le chapitre 5, nous proposons donc des estimateurs de courbes totales ou moyennes de population en présence de valeurs manquantes. Afin de limiter la détérioration de la qualité des estimations dues à ces valeurs manquantes, on propose et on compare différentes approches. On applique notamment des méthodes non paramétriques de lissage de données fonctionnelles adaptées au contexte des sondages et des valeurs manquantes. Une autre approche repose sur l'adaptation au cadre des données fonctionnelles des méthodes d'imputation de valeurs manquantes en sondages telles que les plus proches voisins. Enfin, on propose une variante de la technique de l'interpolation linéaire qui permet de prendre en compte la tendance des autres courbes de l'échantillon. Pour chacun des estimateurs considérés, on propose un estimateur de variance associé, en se basant sur les travaux de Beaumont and Bissonnette (2011) sur l'imputation composite.

Ces différents travaux ont fait l'objet d'articles publiés ou soumis à différentes revues ainsi que de présentations dans des conférences :

— Cardot, H., De Moliner, A., & Goga, C. (2015). Estimating with kernel smoothers the mean of functional data in a finite population setting. A note on variance estimation in presence of partially observed trajectories. Statistics & Probability Letters, 99, 156-166.

— Cardot, H., De Moliner, A., & Goga, C. Estimation of total electricity consumption curves by sampling in a finite population when some trajectories are partially unobserved, soumis à Canadian Journal of Statistics

— Cardot, H., De Moliner, A., & Goga, C. Robust estimation of total electricity

consumption curves by sampling in a finite population (à soumettre prochainement)

— De Moliner A., Estimation de courbes moyennes de consommation électrique à partir d'échantillon pour des petits domaines, soumis à Survey Methodology pour un numéro spécial

— De Moliner, A., Goga, C. and Cardot, H. (2016). Estimation of total electricity consumption curves of small areas by sampling in a finite population. Compstat 2016, Eds Colubi, A., Blanco, A. and Gatu, C., 49-57.

Chapitre 2

État de l'art sur l'estimation par sondage pour des données fonctionnelles

Dans ce Chapitre, nous présentons quelques éléments introductifs concernant l'estimation par sondage pour des données fonctionnelles. Nous nous intéressons tout d'abord à l'analyse des données fonctionnelles : on évoque en particulier la question du passage en dimension finie et de la réduction de dimension, qui comme nous le verrons plus tard sera centrale dans notre démarche d'adaptation des concepts existants en sondage au cadre des données fonctionnelles. On présente également les modèles linéaires qui nous seront utiles dans le Chapitre 4 pour modéliser les relations entre courbes de consommation électrique et variables explicatives afin de réaliser des estimations sur des petites sous-populations.

Ensuite, dans une seconde partie nous introduisons quelques notions de théorie des sondages : on introduit notamment certains des estimateurs les plus couramment utilisés, à savoir l'estimateur de Horvitz-Thompson, et l'estimateur par calage et on aborde la question de l'estimation de variance sous le plan mais aussi de la distinction entre approche basée sur le modèle et approche basée sur le plan. On s'intéresse plus spécifiquement au cas où la variable d'intérêt est une courbe.

Du fait de la richesse et de la variété de la littérature concernée, cet état de l'art est très partiel. Pour de plus amples détails sur les différents points traités, le lecteur pourra se référer par exemple aux ouvrages de Tillé (2001), Ardilly (2006), Särndal et al. (2003) en ce qui concerne l'estimation par sondage et à Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) ou encore Cuevas (2014) pour l'analyse des données fonctionnelles. En outre, au début de chaque Chapitre on proposera quelques références bibliographiques spécifiques à chacun des points abordés.

2.1 Données fonctionnelles

L'analyse des données fonctionnelles est une branche de la statistique s'intéressant à l'étude de données de types courbes ou fonctions. Elle trouve des applications dans des domaines variés tels que la médecine (analyse de données d'électroencéphalogrammes par exemple), l'économie ou encore l'industrie (avec notamment la maintenance prédictive basée sur l'analyse de données de capteurs).

Les premiers outils d'analyse de données fonctionnelles sont apparus dans les an-

nées 1970 (voir par exemple Deville (1974) ou Dauxois et al. (1982) mais se sont davantage développés dans les années 1990 et 2000 avec l'essor des capacités de traitement et de stockage des données et surtout la multiplication des données issues de capteurs dans différents domaines scientifiques et techniques. Les courbes de consommation électrique remontées par des compteurs communicants sont d'ailleurs un exemple typique de cette catégorie de données.

Le fait de traiter les courbes en tant que telles et non pas simplement par des méthodes de statistique multivariée a de nombreux avantages : cela permet d'exploiter les régularités des courbes et leur structure temporelle intrinsèque afin que les estimations des valeurs des courbes moyennes aux différents instants se consolident entre elles. De plus, cela permet d'éviter les problèmes de colinéarité que l'on rencontrerait en utilisant des méthodes multivariées sur les courbes discrétisées.

Les principaux ouvrages de référence sur les données fonctionnelles sont Ramsay and Silverman (2005), Horváth and Kokoszka (2012) et Ferraty and Vieu (2006).

Dans la suite de cette section, largement basée sur Ramsay and Silverman (2005), après avoir introduit quelques notations et explicité les spécificités du contexte des travaux EDF, on présente les principaux outils d'analyse des données fonctionnelles que nous utilisons dans le cadre de cette thèse pour adapter les méthodes usuelles de sondage à notre problématique : nous évoquons d'abord la question du passage en dimension finie avec notamment la projection des courbes sur des bases d'ondelettes, mais aussi l'Analyse en Composantes Principales Fonctionnelle qui est une méthode de réduction de dimension et enfin les modèles linéaires fonctionnels.

2.1.1 Cadre de travail

Hypothèses et notations sur les données fonctionnelles

Soit une population d'intérêt U, composée de N éléments :

$$\mathrm{U} = \{1, \dots, i, \dots, \mathrm{N}\}.$$

Soit Y une variable d'intérêt définie pour chaque individu i de la population. Dans toute la suite de ces travaux, la variable d'intérêt Y considérée n'est pas un réel comme dans le cas le plus usuel traité dans le contexte des sondages mais une courbe définie sur un intervalle de temps $[0, \mathrm{T}]$. On suppose que Y appartient à $\mathrm{L}^2[0, \mathrm{T}]$, l'espace de Hilbert des fonctions définies sur $[0, \mathrm{T}]$ de carré intégrable.

A chaque unité i de notre population U de taille N, on associe donc une courbe, ou trajectoire, c'est-à-dire une fonction $\mathrm{Y}_i(t)$, $t \in [0, \mathrm{T}]$, où l'indice continu t représente le temps.

On note Y_i la valeur prise par Y pour l'individu i. On souhaite estimer une fonction de la variable d'intérêt

$$\theta = \theta(\mathrm{Y}_i, i \in \mathrm{U}),$$

que l'on appelle *paramètre d'intérêt*. On s'intéresse en particulier fréquemment à deux paramètres :

— la courbe totale de U à chaque instant :

$$t_{\mathrm{Y}}(t) = \sum_{i \in \mathrm{U}} \mathrm{Y}_i(t), \quad t \in [0, \mathrm{T}]$$

— la courbe moyenne de U à chaque instant :

$$\mu_Y(t) = \frac{1}{N}\sum_{i\in U} Y_i(t), \quad t \in [0,T].$$

En pratique, les courbes ne sont pas observées continûment pour tout $t \in [0,T]$ mais seulement sur un ensemble fini L d'instants de mesure $0 = t_1 < \ldots < t_l < \ldots < t_L = T$, que l'on suppose par la suite identiques pour l'ensemble des unités et équidistants. On doit donc se poser la question de la reconstruction de la forme fonctionnelle des courbes à partir de ces données discrétisées. Pour cela, une stratégie usuelle consiste à réaliser des interpolations ou encore des lissages, selon des techniques décrites par exemple dans Ramsay and Silverman (2005). Ces méthodes ont été utilisées dans le cadre des sondages dans Cardot and Josserand (2011) pour l'interpolation linéaire et Cardot et al. (2013a) pour le lissage. Dans ce contexte, si le nombre de points de discrétisation est suffisant et les trajectoires sont assez régulières, alors l'erreur d'approximation due au lissage ou à l'interpolation est négligeable par rapport à l'erreur d'échantillonnage.

Enfin, on suppose qu'il n'y a pas de valeurs manquantes, sauf dans le Chapitre 5 qui traite justement de cette problématique.

Par ailleurs, on considère également un vecteur $\mathbf{X}_i = (X_{i1},\ldots,X_{ip})' \in \mathbb{R}^p$ composé de p variables explicatives mesurées sur chaque individu i de la population et liées à notre variable d'intérêt Y (la courbe de consommation électrique). On remarque bien qu'on considère ici des variables $\mathbf{X}_i$ **scalaires et non fonctionnelles**.

Application au contexte EDF

Dans le contexte EDF, les courbes considérées sont bien évidemment des courbes de consommation électrique, aussi appelées courbes de charge. Les courbes de charge totales et par abus de langage les courbes de charge moyennes sont aussi appelées "synchrones de consommation", ou tout simplement "synchrones". De plus, la population U considérée est souvent une catégorie de clients présentant des caractéristiques communes (ensemble des clients résidentiels, ensemble des clients possédant un chauffage électrique, ...).

Les valeurs des consommations électriques moyennes d'un groupe de clients à différents instants présentent bien évidemment des corrélations très fortes et de structures complexes (saisonnalités annuelles, hebdomadaires et infrajournalières, dépendance plus ou moins importantes aux températures extérieures,...) c'est pourquoi l'utilisation de méthodes d'analyse des données fonctionnelles apparaît ici judicieuse pour exploiter les fortes dépendances temporelles de notre problématique afin de tenter de gagner en précision.

On remarque toutefois que, dans le cadre de certains cas d'applications tels que les mécanismes réglementaires, on souhaite en réalité estimer la valeur de la courbe moyenne non pas en chaque instant de l'intervalle $[0,T]$ mais uniquement en chacun des instants de discrétisation $0 = t_1 < \ldots < t_l < \ldots < t_L = T$. On peut alors considérer que, en termes de finalité, notre problème relève plutôt de la statistique en grande dimension que de l'analyse des données fonctionnelles. L'aspect fonctionnel de notre

problématique est dans ce cas un moyen et non une fin en soi (*i.e.* on utilise des méthodes issues de l'analyse des données fonctionnelles pour améliorer l'estimation de nos quantités d'intérêt que sont les vecteurs des valeurs des courbes moyennes ou totales de la population aux instants de discrétisation).

Enfin, les courbes de consommation électrique que nous étudions sont souvent "chahutées", c'est pourquoi lorsque l'on souhaitera projeter des courbes dans un espace de dimension finie, on privilégiera l'utilisation de bases d'ondelettes à d'autres comme les bases de Fourier ou des B-splines par exemple.

Pour illustrer nos propos, voici quelques exemples de courbes de consommation électrique de clients résidentiels irlandais. Ces courbes sont extraites d'un jeu de données proposé en accès libre par la commission irlandaise pour la régulation de l'énergie (CER) et que nous présentons plus en détail dans le Chapitre 3. Ces courbes ont été mesurées au pas demi-horaire pendant une semaine. Elles comportent donc 336 points chacune. On observe nettement les saisonnalités infrajournalières des consommations (les ménages consomment davantage le jour que la nuit) et on remarque que ces courbes sont très irrégulières. Enfin, il arrive que les courbes de certains clients soient nulles la majeure partie du temps, ce qui pourrait correspondre par exemple à des résidences secondaires.

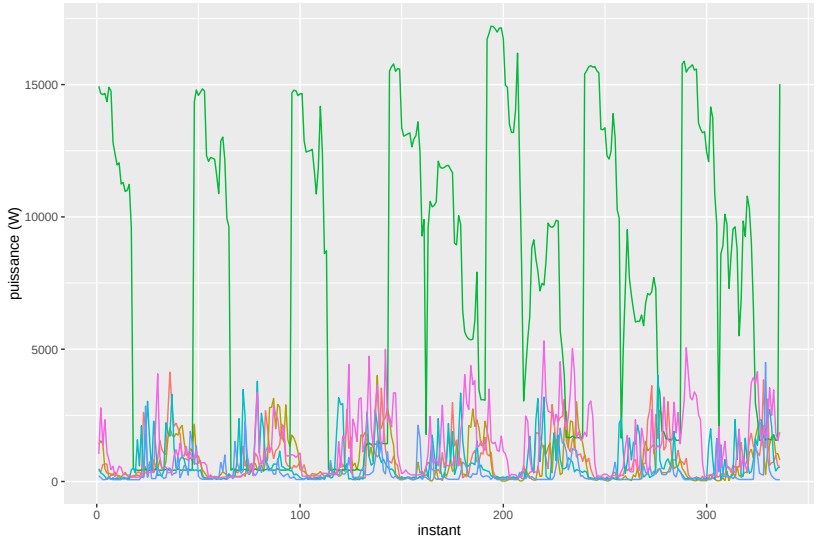

FIGURE 2.1 – Courbes de consommation électrique de quelques clients résidentiels sélectionnés au hasard, au pas demi horaire pendant une semaine (336 points)

En outre, dans le contexte de l'étude des courbes de consommation électrique, que ce soit pour le distributeur ou le commercialisateur, on dispose fréquemment de variables explicatives se présentant sous la forme de données issues de la facturation telles que la consommation annuelle totale d'électricité de chaque individu, le tarif, les options tarifaires et la puissance souscrite, ainsi que les plages d'heures creuses le cas échéant.[1] Ces données sont en général connues pour chaque unité de la

1. Dans les tarifs à heures creuses, l'électricité est moins chère à certaines heures de la journée, ce

population. On connaît également la localisation géographique du client, qui influe sur sa consommation, du fait par exemple du climat mais aussi du type rural ou urbain de sa commune. Cette localisation peut aussi nous permettre de récupérer des données ouvertes disponibles sur un ensemble plus grand d'individus, par exemple les données fournies par l'INSEE à la maille de zones géographiques fines d'environ 1500 habitants appelées IRIS et constituant une source précieuse d'information sur les caractéristiques socio-démographiques des ménages ainsi que sur certaines caractéristiques des logements (taux de résidences secondaires ou d'appartements par exemple).

Certaines informations sont disponibles avant la sélection de l'échantillon tandis que d'autres ne peuvent l'être qu'a posteriori. Ainsi, que l'on se place dans un contexte futur après le déploiement des compteurs communicants ou dans le contexte actuel, on connaît fréquemment la consommation globale du client sur une période temporelle agrégée incluant la période d'étude : par exemple on peut connaître la consommation mensuelle de chaque client et souhaiter estimer la courbe moyenne de la population pour chaque demi heure du mois considéré.

En outre, certaines variables telles que les équipements électriques, l'année de construction du logement et la composition des familles ne peuvent être connues qu'en soumettant un questionnaire aux individus concernés. La passation de questionnaire est coûteuse mais reste nécessaire dans le cadre de certaines études par exemple si l'son veut étudier l'impact d'un équipement spécifique, tel que les pompes à chaleur, sur la courbe de consommation électrique.

2.1.2 Projection de courbes en dimension finie

La principale difficulté soulevée par l'analyse des données fonctionnelles réside dans le fait que l'on travaille sur des objets appartenant à un espace de dimension infinie. Une des premières étapes des méthodes d'analyse des données fonctionnelles va donc consister à se ramener à un espace de dimension finie, plus simple à appréhender et dans lequel on peut appliquer les outils classiques de statistique multivariée. Comme les instants de mesure sont identiques pour l'ensemble des unités de la population, la solution la plus simple pour cela est de travailler sur les vecteurs des L variables aléatoires correspondant aux valeurs de la courbe aux différents instants de discrétisation $\mathbf{Y}_i' = (\mathrm{Y}_i(t_1), \ldots, \mathrm{Y}_i(t_l), \ldots, \mathrm{Y}_i(t_\mathrm{L})$. Cependant, nous verrons que d'autres projections peuvent être plus pertinentes pour traduire les corrélations intertemporelles des consommations électriques.

Les techniques usuelles d'analyse de données fonctionnelles consistent en effet à projeter les courbes sur une base adéquate. Cela permet non seulement de se ramener à un problème de dimension finie mais également de réduire le bruit ou les erreurs de mesure potentielles en ne gardant que les caractéristiques les plus importantes de la courbe. Plus le nombre de fonctions de base est élevé, plus on se rapproche de la courbe mesurée mais plus on risque de conserver du bruit.

qui a un impact sur la répartition infrajournalière des consommations des clients. Ces tarifs ont été conçus pour fluidifier le réseau électrique en reportant une partie des usages électriques, par exemple les chauffe-eau, sur des plages horaires moins chargées.

On considère donc une base $\Phi = \{\phi_1, \ldots, \phi_q \ldots, \phi_Q\}$ de fonctions de $L^2([0,T])$. Une courbe Y peut être approximée par une combinaison linéaire des vecteurs de la base :

$$Y_i(t) = \sum_{q=1}^{Q} \alpha_{iq}\phi_q(t) + \epsilon_i(t), \tag{2.1}$$

avec $Q \in \mathbb{N}$ le nombre de vecteurs de base, $\epsilon_i(t)$ la différence entre la courbe et sa projection et enfin $\alpha_{iq} \in \mathbb{R}$ le cœfficient de l'unité i pour le q-ème vecteur de la base.

Pour projeter les courbes dans la base sélectionnée, on minimise un critère de moindres carrés sur l'erreur de projection auquel on pourra également ajouter un terme de pénalité destiné à favoriser les fonctions les plus lisses, c'est-à-dire dont la norme L^2 de la dérivée d'un certain ordre est la plus faible possible. En effet, le fait d'utiliser des fonctions plus lisses permet de mieux exploiter les régularités de la courbe, donc de réduire la variance au prix d'un peu de biais (la courbe est moins bien ajustée aux données).

Les bases les plus couramment utilisées sont les bases de Fourier pour les fonctions périodiques, les bases de B-splines pour les fonctions non périodiques ou encore les ondelettes, qui comme nous le verrons un peu plus bas sont particulièrement adaptées aux courbes irrégulières et apparaissent de ce fait pertinentes lorsque l'on travaille sur des courbes de consommation électrique, très chahutées (voir par exemple Antoniadis et al. (2013)). Dans la suite, nous nous concentrons sur ces dernières.

Projection sur une base d'ondelettes Les bases d'ondelettes permettent de combiner l'aspect "multi-échelle" des séries de Fourier et le caractère temporellement localisé des splines. Elles permettent d'approximer des fonctions sur $(-\infty, \infty)$ de carré intégrable. Pour plus de détails, on pourra se référer à Mallat (1999). Le principe de la décomposition en ondelettes est de choisir une fonction mère ψ et de considérer l'ensemble des dilatations et translations de cette fonction de la forme

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k),$$

pour des entiers j et k. L'ondelette mère ψ est choisie de façon à ce que la base soit orthogonale. L'intérêt de cette décomposition est qu'elle permet une analyse "multi-échelle" du fait des dilations et permet de capter des spécificités localisées temporellement grâce aux translations. Cela lui permet donc d'être adaptée aux fonctions discontinues ou présentant des changements rapides. En pratique, l'algorithme DWT (Discrete Wavelet Transform) permet d'approximer rapidement les cœfficients d'ondelettes.

2.1.3 Réduction de dimension : Analyse en Composantes Principales fonctionnelle

Tout comme l'ACP (Analyse en Composantes Principales) en dimension finie, l'ACP fonctionnelle est une méthode de réduction de dimension permettant de résumer l'information contenue dans un jeu de données. Elle a été proposée par Deville (1974), ses propriétés théoriques ont été étudiées dans Dauxois et al. (1982) ou Hall et al. (2006) et enfin elle a été adaptée dans le cadre des sondages par Cardot et al. (2010).

Concrètement, le principe de l'ACP est de chercher une base orthonormale de vecteurs (ici des fonctions) tels que la projection des données sur celle-ci permette de garder le maximum d'information (de variance). Pour cela, il faut considérer la fonction de variance-covariance dont on trouve les vecteurs propres qui constitueront les vecteurs de base.

Plus formellement, on définit la fonction de variance covariance $v(s,t)$ par

$$v(s,t) = \mathrm{N}^{-1} \sum_{i=1}^{\mathrm{N}} (\mathrm{Y}_i(s) - \mu_{\mathrm{Y}}(s))(\mathrm{Y}_i(t) - \mu_{\mathrm{Y}}(t)) \quad \forall s,t \in [0,\mathrm{T}]. \tag{2.2}$$

On définit également l'*opérateur de covariance* $\mathcal{V}$ sur $\mathrm{L}^2([0,\mathrm{T}])$ par

$$f \to \mathrm{V}f = \int_0^{\mathrm{T}} v(.,t) f(t) dt. \tag{2.3}$$

On cherche ensuite les valeurs propres $(\lambda_k)_{k\geq 1}$ et vecteurs propres orthonormés associés $(\zeta_k)_{k\geq 1}$ de cet opérateur, tels que

$$\mathrm{V}\zeta_k = \lambda_k \zeta_k,$$

avec $\lambda_1 \geq \lambda_2 \ldots \geq 0$ et $\int_0^{\mathrm{T}} \xi_k(t)\zeta_{k'}(t)dt = 1$ si $k = k'$ et 0 sinon.

On peut ensuite écrire chaque courbe sous la forme de l'expansion de Karhunen-Loève (voir par exemple Ramsay and Silverman (2005)) :

$$\mathrm{Y}_i(t) = \mu_{\mathrm{Y}}(t) + \sum_{k=1}^{\mathrm{K}} g_{ik}\zeta_k(t) + \mathrm{R}_i(t), \tag{2.4}$$

avec K le nombre de composantes que l'on choisit de garder, μ_{Y} le centre de l'espace (c'est-à-dire la moyenne de Y sur la population), g_{ik} le score de l'unité i pour la composante k et $\mathrm{R}_i(t)$ un terme d'erreur représentant l'écart entre la courbe i et sa projection. Le score g_{ik} peut être calculé comme le produit scalaire de ζ_k and $\mathrm{Y}_i - \mu_{\mathrm{Y}}$:

$$g_{ik} = \int_0^{\mathrm{T}} (\mathrm{Y}_i - \mu_{\mathrm{Y}})(t)\zeta_k(t)dt.$$

Remarque 1. *Dans le cadre de cette thèse, nous allons utiliser l'ACP fonctionnelle non pas réellement comme un outil de réduction de dimension, mais plutôt comme un outil permettant de transformer un problème fonctionnel donc de dimension infinie en un nombre fini de problèmes univariés décorrélés (ou en tous cas moins corrélés qu'à l'origine) que l'on peut résoudre séparément en utilisant les méthodes d'estimation par sondage de totaux réels. Pour cette raison, on garde en général un nombre de composantes* K *grand, afin de ne pas perdre trop d'information. On peut alors considérer que l'erreur d'approximation est négligeable par rapport à la variance d'échantillonnage. D'un point de vue plus conceptuel, puisqu'on travaille en population finie, le "bruit" des courbes de charge moyennes correspond souvent à une réalité qu'on souhaite mesurer et non pas filtrer : on ne s'intéresse en effet pas à l'espérance* $\mathbb{E}(\mathrm{Y})$ *de notre variable d'intérêt mais bien à la moyenne* μ_{Y} *des réalisations de cette variable sur notre population* U. *Le bruit est donc inclus dans cette moyenne.*

En pratique, comme nous l'avons déjà évoqué, nos données fonctionnelles ne sont observées que pour un ensemble fini d'instants de discrétisation équidistants. La recherche de vecteurs propres ne se fait donc pas dans l'espace des fonctions mais dans un espace de dimension finie, obtenu soit en travaillant sur les courbes discrétisées, soit à l'aide d'une projection sur une base de fonctions telles que les ondelettes par exemple.

Dans le premier cas, comme les instants de mesure sont équidistants, on pourra réaliser une ACP classique sur les valeurs des courbes aux instants de discrétisation $\mathbf{Y}_i^{'} = (\mathrm{Y}_i(t_1), \ldots, \mathrm{Y}_i(t_l), \ldots, \mathrm{Y}_i(t_\mathrm{L}))$.

Dans le second cas, on peut également transformer notre problème fonctionnel en un problème de recherche des vecteurs propres d'une matrice, que l'on sait résoudre (voir Ramsay and Silverman (2005)) : soit $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{i\mathrm{Q}})'$ le vecteur des cœfficients d'expansion pour la courbe Y_i dans la base $\Psi = (\psi_1, \ldots, \psi_\mathrm{Q})'$, tel que

$$\mathrm{Y}_i(t) \approx \Psi(t)'\alpha_i,$$

où $\Psi(t) = (\psi_1(t), \ldots, \psi_\mathrm{Q}(t))'$. Soit $\mathbf{A}$ la matrice $\mathrm{N} \times \mathrm{Q}$ dont les lignes sont les vecteurs $\alpha_i^{'}$ et $\mathbf{W} = \int_0^\mathrm{T} \Psi_t \Psi_{t'} dt$ la matrice symétrique de taille $\mathrm{Q} \times \mathrm{Q}$ des produits scalaires entre les fonctions de base. Notre problème se ramène finalement au problème matriciel classique de la recherche des vecteurs $\mathbf{u}$ tels que

$$\mathrm{N}^{-1}\mathbf{W}^{-1/2}\mathbf{A}'\mathbf{A}\mathbf{W}^{-1/2}\mathbf{u} = \lambda\mathbf{u}. \tag{2.5}$$

On peut ensuite en déduire les vecteurs propres recherchés dans l'espace de projection $\mathbf{b} = \mathbf{W}^{-1/2}\mathbf{u}$.

Remarque 2. *Lorsque la base est orthonormale, ce qui est le cas pour les bases de Fourier ou certaines familles d'ondelettes comme les ondelettes de Daubechies par exemple, on a* $\boldsymbol{W} = \boldsymbol{I}$ *(où* $\boldsymbol{I}$ *est la matrice Identité) et l'ACP fonctionnelle revient à faire l'ACP multivariée sur les cœfficients de la projection.*

Nos échantillons de courbes de charge sont susceptibles de contenir des unités atypiques, c'est pourquoi dans le cadre de nos travaux on préfère souvent utiliser une version robuste de l'ACP, appelée ACP sphérique, et proposée par Locantore et al. (1999), que nous présentons dans le Chapitre 3.

2.1.4 Régression linéaire pour des données fonctionnelles

Dans le Chapitre 4, nous cherchons à modéliser le lien entre les courbes de consommation électrique et l'information auxiliaire. Pour cela, nous utilisons entre autres des modèles de régression linéaire fonctionnelle.

Nous nous proposons donc dans cette sous-section d'introduire quelques éléments sur ces modèles, hors du contexte des sondages. Il existe deux types d'approches en régression fonctionnelle : les approches non paramétriques, telles que celles proposées dans Ferraty and Vieu (2006) qui se basent sur des estimateurs à noyau de type Nadaraya-Watson et que nous ne développons pas ici, et les approches paramétriques, décrites dans Ramsay and Silverman (2005). On utilise en outre différentes méthodes d'estimation de modèles linéaires fonctionnels, selon que la variable à prédire et/ou

les variables explicatives sont fonctionnelles. Dans cette thèse, on se limite au cas où **la variable à prédire est fonctionnelle et on dispose de plusieurs variables explicatives scalaires**.

Nous allons donc postuler une relation linéaire entre les variables explicatives et notre variable d'intérêt. Plus précisément, ce modèle s'écrit alors (voir le Chapitre 13 de Ramsay and Silverman (2005) ou encore dans Faraway (1997)) :

$$Y_i(t) = \mathbf{X}_i^{'}\boldsymbol{\beta}(t) + \epsilon_i(t), \quad \forall i \in U. \tag{2.6}$$

Pour estimer ce modèle, la solution proposée se base sur la projection des courbes en dimension finie (soit en utilisant le vecteur des valeurs aux instants de discrétisation, soit par une projection sur une base adaptée, par exemple une base d'ondelettes) puis l'utilisation dans cette base de l'algorithme classique de régression linéaire, pour une réponse multivariée. Ce modèle a été adapté au contexte des estimations en population finie par Cardot et al. (2013b).

2.2 Sondages

Dans cette partie, on présente quelques notions et on introduit quelques notations sur l'estimation d'un total ou moyenne par sondage en population finie, plus particulièrement dans le cas où la variable d'intérêt est une courbe.

2.2.1 Notations sur les sondages

Comme évoqué dans la section précédente, on souhaite estimer un paramètre θ (bien souvent la moyenne ou le total) dépendant des valeurs de la variable Y sur notre population d'intérêt. Une solution pour cela serait de collecter l'ensemble des valeurs de Y_i pour chacune des unités i de la population afin ensuite d'en déduire la quantité qui nous intéresse. Cependant, pour des raisons de coût ou de rapidité, il est fréquent que l'on ait recours à des enquêtes par sondage, que ce soit dans le domaine de la statistique publique ou dans le privé.

Ainsi, dans le contexte de EDF, bien que des compteurs communicants soient actuellement en cours de déploiement chez l'ensemble des clients professionnels et résidentiels français, on préférera continuer à travailler sur des échantillons de courbes de consommation électrique sélectionnées selon un plan de sondage aléatoire. En effet, la remontée automatique de l'ensemble des courbes de consommation au pas dix minutes collectées par ces compteurs représenterait un coût très important. De plus, l'utilisation par un fournisseur ou distributeur d'électricité de ces données est soumise à l'accord explicite du client du fait de la nature personnelle de cette information, ce qui empêche la collecte systématique. Enfin, dans un contexte d'études de marketing et de connaissance de la clientèle, il est nécessaire de collecter des informations relatives aux caractéristiques socio-démographiques ainsi qu'aux équipements électriques et aux caractéristiques du bâti pour les clients concernés afin d'analyser

l'impact de ces éléments sur les courbes de charges associées. Cela rend nécessaire la passation de questionnaires et donc la collecte par sondage.

On suppose donc qu'on dispose d'un échantillon s de taille n_s, sélectionné dans U selon un plan de sondage aléatoire p, sans remise, où $p(\cdot)$ est une loi de probabilité sur l'ensemble des parties $\mathscr{S}$ de U. Le plan de sondage p vérifie

$$\forall s \in \mathscr{S}, p(s) \geq 0$$

et

$$\sum_{s \in \mathscr{S}} p(s) = 1.$$

Dans la suite, on suppose de plus que le plan est de taille fixe, notée n et qu'il ne dépend pas du temps t.

Plaçons-nous pour l'instant dans l'approche basée sur le plan, aussi appelée *approche design-based* : le seul aléa considéré ici est le fait qu'un individu appartienne ou non à l'échantillon et les valeurs Y_i sont considérées comme déterministes. On note $\mathbb{E}_p(\cdot)$ (respectivement $\mathbb{V}_p(\cdot)$) l'espérance (respectivement la variance) d'un estimateur $\hat{\theta}$ par rapport au plan p :

$$\mathbb{E}_p(\hat{\theta}) = \sum_{s \in \mathscr{S}} p(s)\hat{\theta}_s,$$

où $\hat{\theta}_s$ est la valeur de $\hat{\theta}$ sur l'échantillon s et

$$\mathbb{V}_p(\hat{\theta}) = \sum_{s \in \mathscr{S}} p(s)(\hat{\theta}_s - \mathbb{E}_p(\hat{\theta}))^2.$$

Dans cette approche, le biais d'un estimateur, qui représente l'erreur commise en moyenne, se définit par

$$\mathrm{B}_p(\hat{\theta}) = \mathbb{E}_p[\hat{\theta} - \theta] = \sum_{s \in \mathscr{S}} p(s)[\hat{\theta}_s - \theta].$$

Dans notre contexte où le paramètre d'intérêt θ est une courbe et non un réel, on définit une fonction de variance ainsi qu'une fonction de covariance sur $[0, \mathrm{T}]$:

$$\mathbb{V}_p(\hat{\theta}(t)) = \sum_{s \in \mathscr{S}} p(s)(\hat{\theta}_s(t) - \mathbb{E}_p(\hat{\theta}(t)))^2 \quad \forall t \in [0, \mathrm{T}]$$

et

$$Cov_p(\hat{\theta}(t), \hat{\theta}(u)) = \sum_{s \in \mathscr{S}} p(s) \left\{\hat{\theta}_s(t) - \mathbb{E}_p(\hat{\theta}(t))\right\} \left\{\hat{\theta}_s(u) - \mathbb{E}_p(\hat{\theta}(u))\right\} \quad \forall t, u \in [0, \mathrm{T}].$$

On cherche fréquemment à proposer des estimateurs non biaisés, c'est-à dire de biais nuls, et dont la variance est la plus faible possible. Cependant, dans le cadre de l'estimation robuste, nous verrons qu'il est parfois pertinent d'utiliser des estimateurs biaisés qui en contrepartie ont une variance plus faible. Ce compromis entre biais et variance peut se formaliser à l'aide d'une mesure globale de précision, appelée *Erreur Quadratique Moyenne (EQM)*. Pour une courbe, l'erreur quadratique en un instant t est définie par

$$\mathrm{EQM}(\hat{\theta}(t)) = \mathbb{E}_p\left[(\hat{\theta}(t) - \theta(t))^2\right] = \mathrm{B}_p[\hat{\theta}(t)]^2 + \mathbb{V}_p[\hat{\theta}(t)] \quad \forall t \in [0, \mathrm{T}]. \tag{2.7}$$

Exemple du sondage aléatoire simple sans remise Le plan de sondage le plus basique est le sondage aléatoire simple sans remise. Il présente l'avantage de ne pas nécessiter d'information auxiliaire pour être mis en œuvre, ce qui peut est particulièrement pertinent lorsqu'aucune information n'est disponible sur la population étudiée ou encore si on n'a pas d'idée a priori sur les variables explicatives connues qui influent sur nos variables d'intérêt. En outre, le sondage aléatoire stratifié que nous présentons ci-dessous en est une amélioration.

Le sondage aléatoire simple sans remise consiste à tirer aléatoirement sans remise un échantillon de taille n fixée parmi les N individus de la population, de sorte que chacun des échantillons possibles de n individus distincts ait la même probabilité de sélection $p(s)$. Celle-ci est donc égale à l'inverse du nombre d'échantillons distincts de taille n pouvant être tiré dans une population de taille N. Par conséquent, chaque unité aura également la même probabilité d'être tirée. On a donc :

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si} \quad \#(s) = n \\ 0 & \text{sinon} \end{cases}$$

Nous allons maintenant présenter un estimateur couramment utilisé en sondages : l'estimateur de Horvitz-Thompson.

2.2.2 Estimateur de Horvitz-Thompson

On note $\mathbb{1}_i = \mathbb{1}_{\{i \in s\}}$ l'indicatrice d'appartenance de l'individu i à l'échantillon s, qui vaut 1 si l'individu appartient à s et 0 sinon. Dans l'approche basée sur le plan cette quantité est aléatoire.

Pour un plan de sondage p, on note π_i la *probabilité d'inclusion d'ordre 1* de l'unité i c'est-à-dire la probabilité que l'unité i soit sélectionnée dans un échantillon,

$$\pi_i = \mathrm{P}r(i \in s) = \mathrm{P}r(\mathbb{1}_i = 1) = \sum_{s\, tq\, i \in s} p(s), \quad \forall i \in \mathrm{U}.$$

De même, on note π_{ij} la probabilité d'inclusion d'ordre 2 des unités i et j c'est-à-dire la probabilité que les unités i et j soient sélectionnées dans un même échantillon,

$$\pi_{ij} = \mathrm{P}r(i \in s\, et\, j \in s) = \mathrm{P}r(\mathbb{1}_i \mathbb{1}_j = 1) = \sum_{s, \quad tq \quad i,j \in s} p(s), \quad \forall i, j \in \mathrm{U}.$$

On suppose que les probabilités d'ordre 1 et 2 sont strictement positives pour tout $i, j \in \mathrm{U}$. Dans notre contexte fonctionnel où l'on s'intéresse à des courbes, ces probabilités restent cependant constantes au cours du temps. Pour une étude du cas où les probabilités d'inclusion varient au cours du temps, on pourra se référer à Degras (2014).

Pour un plan de sondage $p(\cdot)$ fixé, la fonction $\mathbb{1}_i$ a les propriétés suivantes (voir par exemple Tillé (2001), Chapitre 3) :

Résultat :
Soit un plan de sondage p. Alors pour tout i et j dans U, on a :

1. $\mathbb{E}_p(\mathbb{1}_i) = \pi_i$;

2. $\mathbb{V}_p(\mathbb{1}_i) = \pi_i(1-\pi_i) = \Delta_{ii}$;
3. $Cov_p(\mathbb{1}_i, \mathbb{1}_j) = \pi_{ij} - \pi_i\pi_j = \Delta_{ij}, i \neq j$.

Etant donné un plan de sondage $p(.)$, Horvitz et Thompson (1952) et Narain (1951) ont proposé un estimateur linéaire et sans biais d'un total t_Y valable pour tout plan de sondage

$$\hat{t}_Y^{HT} = \sum_{i\in s} \frac{Y_i}{\pi_i} = \sum_{i\in U} \frac{Y_i}{\pi_i} \mathbb{1}_{\{i\in s\}}.$$

Cet estimateur est aussi appelé *estimateur de Narain-Horvitz-Thompson* ou π-*estimateur* du total t_Y ou encore *estimateur par les valeurs dilatées*. En effet, il affecte à chaque unité échantillonnée un poids égal à l'inverse de sa probabilité d'inclusion dans l'échantillon $d_i = \frac{1}{\pi_i} \geq 1$.

Théorème 1. *Horvitz-Thompson (1952)*
Si pour toute unité $i \in U$ on a $\pi_i > 0$ alors $\hat{t}_{Y\pi}$ est un estimateur sans biais de t_Y.

Lorsque la taille de la population N est connue, on peut estimer la moyenne μ_Y par le π-estimateur en utilisant

$$\hat{\mu}_Y^{HT} = \frac{1}{N} \sum_{i\in s} \frac{Y_i}{\pi_i}. \tag{2.8}$$

Théorème 2. *Horvitz-Thompson (1952)*
Soit $\hat{t}_Y^{HT}$ l'estimateur de Horvitz-Thompson d'un total t_Y. Si $\pi_i > 0$ pour tout $i \in U$, alors il a pour variance

$$\mathbb{V}_p(\hat{t}_Y^{HT}) = \sum_{i\in U} \sum_{j\in U} \left(\pi_{ij} - \pi_i\pi_j\right) \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j}. \tag{2.9}$$

Si $\pi_{ij} > 0 \, \forall i, j \in U$, alors l'estimateur non biaisé de cette variance est donné par

$$\hat{\mathbb{V}}_p(\hat{t}_Y^{HT}) = \sum_{i\in s} \sum_{j\in s} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j}. \tag{2.10}$$

Exemple du sondage aléatoire simple Pour un sondage aléatoire simple sans remise, les probabilités d'inclusion simples et doubles de chaque unité ou couples d'unités de la population sont identiques pour l'ensemble des unités et égales respectivement à

$$\pi_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \quad \forall i \in U,$$
$$\pi_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}, \quad \forall i \neq j \in U.$$

Pour un sondage aléatoire simple, l'estimateur de Horvitz-Thompson de la moyenne est

$$\hat{\mu}_Y^{HT} = \frac{1}{N} \sum_{i\in s} \frac{Y_i}{\pi_i} = \frac{1}{n} \sum_{i\in s} Y_i. \tag{2.11}$$

Sa variance est :

$$\mathbb{V}_p(\hat{\mu}_Y^{HT}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y,U}^2, \tag{2.12}$$

avec $S_{Y,U}^2 = \frac{1}{N-1}\sum_{k\in U}(Y_k - \mu_Y)^2$ la dispersion de la variable d'intérêt Y dans la population. Elle peut être estimée par

$$\widehat{\mathbb{V}}_p(\hat{\mu}_Y^{HT}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{Y,s}^2, \tag{2.13}$$

avec $s_{Y,s}^2 = \frac{1}{n-1}\sum_{i\in s}(Y_i - \hat{\mu}_Y)^2$ la variance corrigée de la variable d'intérêt Y dans l'échantillon et $\hat{\mu}_Y = \frac{1}{n}\sum_{i\in s} Y_i$.

2.2.3 Estimateur de Horvitz-Thompson pour des données fonctionnelles

Dans notre contexte, la variable Y dont on cherche à estimer le total ou la moyenne sur la population U est une courbe, et plus précisément une courbe de consommation électrique. Ce cas a déjà été étudié dans la littérature. Ainsi, Cardot et al. (2010) propose un estimateur de Horvitz-Thompson de la courbe moyenne :

$$\hat{\mu}_Y^{HT}(t) = \frac{1}{N}\sum_{i\in s}\frac{Y_i(t)}{\pi_i} = \sum_{i\in U}\frac{Y_i(t)}{\pi_i}\mathbb{1}_{i\in s}, \quad t \in [0,T]. \tag{2.14}$$

Pour chaque instant t, cet estimateur $\hat{\mu}_Y^{HT}(t)$ est sans biais par rapport au plan :

$$\mathbb{E}_p(\hat{\mu}_Y^{HT}(t)) = \mu_Y(t).$$

Les propriétés de convergence uniforme de cet estimateur ont été démontrées dans Cardot et al. (2013c) et Cardot and Josserand (2011). Ces estimateurs ont été adaptés au cas de données bruitées par Cardot et al. (2013a).

On peut en outre définir la fonction de covariance de type Horvitz-Thompson de l'estimateur de la courbe totale à différents instants :

$$Cov_p\left(\hat{t}_Y^{HT}(t), \hat{t}_Y^{HT}(u)\right) = \sum_{i\in U}\sum_{j\in U}(\pi_j\pi_{ij} - \pi_i\pi_j)\frac{Y_i(t)}{\pi_i}\frac{Y_j(u)}{\pi_j} \quad \forall t,u \in [0,T]. \tag{2.15}$$

Si on suppose comme précédemment que $\pi_{ij} > 0$ pour tous $i, j \in U$, alors un estimateur sans biais de cette covariance est donné par :

$$\widehat{Cov}_p\left(\hat{t}_Y^{HT}(t), \hat{t}_Y^{HT}(u)\right) = \sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\frac{Y_i(t)}{\pi_i}\frac{Y_j(u)}{\pi_j} \quad \forall t,u \in [0,T]. \tag{2.16}$$

En particulier, pour $t = u$, on a l'expression de l'estimateur de variance à un instant donné,

$$\widehat{\mathbb{V}}_p\left(\hat{t}_Y^{HT}(t)\right) = \sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\frac{Y_i(t)}{\pi_i}\frac{Y_j(t)}{\pi_j}, \quad \forall t \in [0,T]. \tag{2.17}$$

Cet estimateur de variance est un estimateur *ponctuel*, c'est-à-dire l'estimateur de la variance de l'estimateur de Horvitz-Thompson pour un instant donné. En suivant Cardot et al. (2013c), il est également possible de construire ce qu'on appelle des *bandes de confiance*, c'est-à-dire des bandes dans laquelle la véritable courbe totale est intégralement contenue avec une probabilité de $(1-\alpha)$ %, pour un risque α fixé. Pour ce faire, les auteurs proposent de déterminer la loi du supremum d'un processus gaussien centré dont la fonction de covariance est la fonction de covariance estimée sur l'échantillon et ensuite à en déduire les bandes de confiance pour le quantile recherché.

2.2.4 Prise en compte de variables explicatives

En sondages, il arrive fréquemment que l'on dispose de variables explicatives très liées à nos variables d'intérêt et dont on connaît les totaux sur la population d'intérêt ainsi que la valeur pour chaque unité de l'échantillon, voire dans certains cas pour chaque unité de la population. Il est donc pertinent de chercher à se servir de ces informations, aussi nommées dans la suite *information(s) auxiliaire(s)* pour améliorer la qualité des estimations. Les manières d'exploiter ces informations sont évoquées dans les trois paragraphes suivants : on présente d'abord le sondage aléatoire stratifié qui les intègre à l'étape de l'échantillonnage, puis on s'intéresse à l'estimateur par calage qui est une des solutions permettant d'exploiter l'information à l'étape de l'estimation. Ensuite on se focalise plus précisément sur l'estimateur de Hájek qui est un cas particulier du calage dans lequel la taille de la population est prise en compte en tant qu'information auxiliaire. Enfin, on présente l'approche basée sur un modèle en sondages, que nous allons notamment mettre en œuvre dans le Chapitre 4 pour proposer des estimations sur de petites sous-populations.

Au niveau de l'échantillonnage : sondage stratifié

L'information auxiliaire peut être exploitée à l'étape de l'échantillonnage de multiples façons : sondages stratifiés, équilibrés, à probabilités inégales notamment. Pour plus de détails sur ces plans de sondage on pourra se référer à Tillé (2001) ou encore Ardilly (2006). Ici on s'intéresse plus particulièrement au sondage stratifié.

Dans le paragraphe précédent, nous avons vu avec l'équation (2.12) que pour le plan de sondage aléatoire simple la variance de l'estimateur de Horvitz-Thompson dépendait de la variance de la variable d'intérêt dans la population. Cet estimateur peut donc être très imprécis si la dispersion de la variable d'intérêt est très élevée dans la population étudiée. Afin de réduire la variance de l'estimateur de Horvitz-Thompson, l'idée de la stratification consiste donc à utiliser l'information auxiliaire pour constituer des groupes d'individus homogènes, dans lequel la variabilité de la variable d'intérêt est faible.

Le principe du sondage stratifié est de partitionner la population en groupes homogènes appelés *strates*, constitués selon notre information auxiliaire liée à la variable d'intérêt, puis ensuite de réaliser un sondage indépendant, souvent aléatoire simple sans remise, au sein de chaque strate. La diminution de la variance est d'autant plus

importante que les strates sont homogènes par rapport à la variable d'intérêt, c'est-à-dire que la somme des variances au sein de chaque strate appelée *variance intra-strates* est faible et donc que la variabilité entre les strates appelée *variance inter-strates*, est forte, c'est-à-dire que les moyennes de la variable d'intérêt sont différentes d'une strate à une autre.

Plus formellement, on suppose que notre population d'intérêt U est partitionnée en H strates disjointes $\mathrm{U}_1, \ldots, \mathrm{U}_\mathrm{H}$ de tailles $\mathrm{N}_1, \ldots, \mathrm{N}_\mathrm{H}$ connues. Ces strates vérifient donc :

$$\bigcup_{h=1}^{\mathrm{H}} \mathrm{U}_h = \mathrm{U} \text{ et } \mathrm{U}_g \bigcap \mathrm{U}_h = \emptyset \text{ pour } g \neq h, \tag{2.18}$$

et

$$\mathrm{N}_1 + \ldots + \mathrm{N}_\mathrm{H} = \mathrm{N}.$$

Dans chacune de ces strates, on sélectionne indépendamment un échantillon s_h de taille connue n_h avec les n_h vérifiant $n_1 + \ldots + n_\mathrm{H} = n$. L'estimateur de Horvitz-Thompson d'un total est défini par

$$\hat{t}^{\mathrm{HT}}_{strat,\mathrm{Y}}(t) = \sum_{h=1}^{\mathrm{H}} \sum_{i \in s_h} \frac{\mathrm{Y}_i(t)}{\pi_i^h} = \sum_{h=1}^{\mathrm{H}} \hat{t}_{h,\mathrm{Y}}(t), \quad t \in [0, \mathrm{T}],$$

où $\hat{t}_{h,\mathrm{Y}}$ est l'estimateur de Horvitz-Thompson du total de Y dans la strate U_h. L'échantillon s est alors défini comme l'union de l'ensemble des échantillons par strates : $s = \cup_{h=1}^{\mathrm{H}} \mathrm{U}_h$.

Les tirages réalisés dans les différentes strates sont indépendants, on en déduit donc la variance de $\hat{t}^{\mathrm{HT}}_{strat,\mathrm{Y}}$ par

$$\mathbb{V}_p(\hat{t}^{\mathrm{HT}}_{strat,\mathrm{Y}}(t)) = \sum_{h=1}^{\mathrm{H}} \mathbb{V}_p(\hat{t}^{\mathrm{HT}}_{h,\mathrm{Y}}(t)), \quad t \in [0, \mathrm{T}]. \tag{2.19}$$

Dans le cas particulier très courant d'un sondage aléatoire simple sans remise dans chaque strate, on a l'estimateur

$$\hat{t}^{\mathrm{HT}}_{strat,\mathrm{Y}}(t) = \sum_{h=1}^{\mathrm{H}} \mathrm{N}_h \underbrace{\frac{1}{n_h} \sum_{i \in s_h} \mathrm{Y}_i(t)}_{\hat{\mu}_h}, \quad t \in [0, \mathrm{T}]$$

dont la variance est

$$\mathbb{V}_p(\hat{t}^{\mathrm{HT}}_{strat,\mathrm{Y}}(t)) = \sum_{h=1}^{\mathrm{H}} \mathrm{N}_h^2 \left(\frac{1}{n_h} - \frac{1}{\mathrm{N}_h} \right) \mathrm{S}^2_{\mathrm{Y}(t),\mathrm{U}_h}, \quad t \in [0, \mathrm{T}]$$

où $\mathrm{S}^2_{\mathrm{Y}(t),\mathrm{U}_h} = \frac{1}{\mathrm{N}_h - 1} \sum_{i \in \mathrm{U}_h} (\mathrm{Y}_i - \mu_h)^2$.

La répartition de l'effectif de l'échantillon entre les différentes strates, c'est-à-dire le choix des différents n_h peut se faire de différentes manières. Une première méthode

consiste à choisir des effectifs de strate proportionnels à la taille de la strate dans la population : c'est l'allocation proportionnelle. Les n_h sont alors tels que [2] :

$$\frac{n_h}{\mathrm{N}_h} = \frac{n}{\mathrm{N}}, \quad h = 1, \ldots, \mathrm{H}.$$

Cependant, cette première façon d'allouer les n_h n'est pas forcément celle qui conduit aux estimateurs les plus précis. Intuitivement, pour minimiser la variance de l'estimateur, on souhaite en effet plutôt surreprésenter les individus appartenant aux strates dans lesquelles la variabilité est la plus grande. Cette idée a été formalisée dans Neyman (1934) sous la forme d'un problème d'optimisation : à taille totale n et à coûts d'enquête fixés, il s'agit de déterminer les tailles n_h de chacune des strates de l'échantillon de façon à ce que la variance de l'estimateur de Horvitz-Thompson pour le plan stratifié soit minimale. Lorsque le coût d'enquête est le même dans toutes les strates, cela donne le programme d'optimisation suivant :

$$\min_{n_1,\ldots,n_{\mathrm{H}}} \mathbb{V}(t_{strat,\mathrm{Y}}) \quad \text{sous la contrainte} \quad \sum_{h=1}^{\mathrm{H}} n_h = n.$$

On résout ce programme à l'aide d'un multiplicateur de Lagrange (voir par exemple Cochran (1977)),

$$n_h = n \frac{\mathrm{N}_h \mathrm{S}_{\mathrm{Y},\mathrm{U}_h}}{\sum_{j=1}^{\mathrm{H}} \mathrm{N}_j \mathrm{S}_{\mathrm{Y},\mathrm{U}_j}}, \quad h = 1, \ldots, \mathrm{H}. \tag{2.20}$$

En pratique, la variable d'intérêt et a fortiori sa variabilité dans chaque strate est inconnue à l'étape de l'échantillonnage. On remplace donc les variances de la variable d'intérêt dans chaque strate $\mathrm{S}_{\mathrm{Y},\mathrm{U}_h}$ par les variances $\mathrm{S}_{\mathrm{X},\mathrm{U}_h}$ d'une variable auxiliaire réelle X, connue et la plus corrélée possible à notre variable d'intérêt. Dans le contexte EDF, cela peut par exemple être une consommation d'électricité moyenne sur une longue période, par exemple annuelle. La taille optimale de chaque strate selon ce critère devient alors :

$$n_h = n \frac{\mathrm{N}_h \mathrm{S}_{\mathrm{X},\mathrm{U}_h}}{\sum_{j=1}^{\mathrm{H}} \mathrm{N}_j \mathrm{S}_{\mathrm{X},\mathrm{U}_j}}, \quad h = 1, \ldots \mathrm{H}, \tag{2.21}$$

et on parle alors d'allocation X-optimale.

En pratique, il est possible que ce critère aboutisse à des tailles de strates $n_h > \mathrm{N}_h$. Dans ce cas on procède de manière itérative en sélectionnant tous les individus de la strate puis en recalculant les n_h des autres strates suivant le même critère, jusqu'à ce qu'on ait $n_h \leq \mathrm{N}_h \quad \forall h = 1, \ldots, \mathrm{H}$.

Cette méthode d'allocation optimale de Neyman a été étendue au cadre des données fonctionnelles par Cardot and Josserand (2011).

Au niveau de l'estimation : estimateur par calage fonctionnel

Dans ce paragraphe, nous allons voir comment exploiter l'information auxiliaire à l'étape de l'estimation. Il arrive en effet que l'information auxiliaire ne puisse pas être utilisée au moment de l'échantillonnage, soit parce qu'elle n'est pas connue

2. En pratique, les tailles de strates n_h ainsi définies seront rarement entières et il faut donc les arrondir à l'unité

pour chaque unité de la population mais seulement sur l'échantillon ainsi que sous forme de totaux sur l'ensemble de la population, soit parce qu'elle ne devient disponible qu'après la sélection de l'échantillon. L'exploitation de ces variables auxiliaires à l'étape de l'estimation a pour objectif de réduire la variance de l'estimateur, en incorporant une information auxiliaire pertinente car liée à la variable d'intérêt, mais aussi de corriger ou réduire certains défauts des plans de sondage tels que les défauts de couverture (fait que certaines unités de la population ne soient en pratique pas échantillonnables, par exemple parce qu'elles sont absentes de la base de sondage, ou encore, dans le contexte de EDF, parce qu'elles refusent que leurs données soient exploitées) ou encore les biais générés par une non réponse non ignorable.

Différents estimateurs permettent d'exploiter ces informations auxiliaires, notamment l'estimateur par calage ou l'estimateur par post-stratification qui en est un cas particulier. On peut également utiliser l'approche des sondages assistés par un modèle, étudiée notamment par Särndal et al. (2003). En effet, Cardot et al. (2013b) ont proposé un estimateur assisté par un modèle pour l'estimation de la courbe moyenne par sondage en population finie : le modèle considéré est le modèle linéaire fonctionnel présenté dans le paragraphe 2.1.4. Dans ce même article, différentes stratégies d'estimation de la courbe moyenne en présence d'information auxiliaire sont comparées sur un jeu de données de courbes de consommation électrique. Dans cette Section, nous nous concentrons sur l'estimateur par calage, qui sera employé par la suite.

On se place toujours dans le cadre de l'estimation basée sur le plan, c'est-à-dire que les valeurs Y_i de la variable d'intérêt pour l'ensemble des individus de la population sont considérées comme fixes et non aléatoires, et que l'aléa provient du fait que chaque unité soit ou non sélectionnée dans l'échantillon.

Soit $\mathbf{X}_i = (\mathrm{X}_{i1}, \ldots, \mathrm{X}_{ip})'$ le vecteur des valeurs des p variables explicatives de l'individu i de la population. Ici, on suppose connu le total de ce vecteur sur la population $t_{\mathbf{X}} = \sum_{i \in \mathrm{U}} \mathbf{X}_i$, ainsi que la valeur de $\mathbf{X}_i$ sur chaque unité de l'échantillon. En revanche, $\mathbf{X}_i$ n'est pas forcément connu pour chaque unité non échantillonnée de la population.

Dans le contexte EDF, on peut par exemple penser à une indicatrice de possession de chauffage électrique, connue sur l'ensemble d'un panel à l'aide de questionnaires, et dont on connaît la moyenne sur la population par des sources externes.

L'idée du calage est de modifier le plus légèrement possible les poids de sondage de Horvitz-Thompson $d_i = \frac{1}{\pi_i}$ de façon à faire coïncider les totaux des variables explicatives estimés sur l'échantillon avec les totaux connus de ces mêmes variables sur la population. Intuitivement, les variables explicatives étant liées à la valeur d'intérêt, on espère qu'en accentuant la ressemblance entre l'échantillon et la population sur les variables $\mathbf{X}_i$ on accentuera également leur ressemblance sur la variable d'intérêt.

Plus formellement, l'estimateur par calage s'écrit

$$\hat{t}_{\mathrm{Y}}^{cal}(t) = \sum_{i \in s} w_{is} \mathrm{Y}_i(t), \tag{2.22}$$

avec w_{is} des nouveaux poids, les plus proches possibles des poids de sondage $(d_i)_{i=1}^n$ au sens d'une certaine distance ou pseudo-distance $\mathrm{G}(w, d)$ définie par le statisticien :

$$\min_{w_i} \sum_{i \in s} d_i \frac{\mathrm{G}(w_i, d_i)}{q_i}$$

sous la contrainte que les totaux des variables explicatives soient estimés exactement sur l'échantillon :

$$t_{\mathbf{X}} = \sum_{i \in \mathrm{U}} \mathbf{X}_i = \sum_{i \in s} w_{is} \mathbf{X}_i. \tag{2.23}$$

La fonction $\mathrm{G}(w, d)$ est supposée positive, dérivable par rapport à w, strictement convexe et telle que $\mathrm{G}(d, d) = 0$. Il s'agit donc d'un problème de minimisation sous contrainte linéaire résolu en utilisant la méthode des multiplicateurs de Lagrange. Les poids q_i sont des cœfficients de pondération qui permettent de déterminer l'importance de chaque unité dans l'échantillon s. Lorsqu'il n'y a pas de raison d'accorder d'importance à certaines unités plutôt que d'autres, on prendra $q_i = 1, \quad \forall i$.

Différentes distances ou pseudo distances peuvent être choisies, aboutissant à différentes méthodes de calage dont les plus connues sont la méthode linéaire, la méthode exponentielle, la méthode logit et la méthode linéaire tronquée. Lorsqu'il n'y a pas de non réponse, les estimateurs obtenus selon les différentes distances sont asymptotiquement équivalents à celui obtenu pour la distance du khi-deux. Ce dernier estimateur est l'estimateur assisté par un modèle fonctionnel (voir Cardot et al. (2013b)). Ses propriétés asymptotiques ont été étudiées en détail dans Lardin (2012) ainsi que Cardot et al. (2013a) et Cardot et al. (2013c). Pour ce dernier estimateur, l'expression des poids est

$$w_{is} = \frac{1}{\pi_i} - q_i \left(\sum_{j \in s} \frac{\mathbf{X}_j}{\pi_j} - t_{\mathbf{X}} \right) \left(\sum_{j \in s} \frac{\mathbf{X}_j \mathbf{X}_j'}{\pi_j} \right)^{-1} \frac{\mathbf{X}_i}{\pi_i}, \quad i \in s. \tag{2.24}$$

L'estimateur par calage du total s'écrit alors

$$\hat{t}_{\mathrm{Y}}^{cal}(t) = \hat{t}_{\mathrm{Y}}^{\mathrm{HT}}(t) - \left(\sum_{j \in s} \frac{\mathbf{X}_j}{\pi_j} - t_{\mathbf{X}} \right)' \hat{\boldsymbol{\beta}}(t), \quad t \in [0, \mathrm{T}], \tag{2.25}$$

où $\hat{\boldsymbol{\beta}}(t)$ est l'estimateur du cœfficient de régression de $\mathrm{Y}_i(t)$ sur $\mathbf{X}_i$:

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{i \in s} \frac{\mathbf{X}_i q_i \mathbf{X}_i'}{\pi_i} \right)^{-1} \sum_{i \in s} \frac{\mathbf{X}_i q_i \mathrm{Y}_i(t)}{\pi_i}, \quad t \in [0, \mathrm{T}].$$

Intuitivement, on peut se dire que plus les informations auxiliaires $\mathbf{X}_i$ ont un pouvoir explicatif fort sur la variable d'intérêt, plus la variance de l'estimateur sera réduite. Cette intuition est formalisée dans l'expression de la variance approximée (voir Deville and Särndal (1992))

$$\mathrm{A}\mathbb{V}(\hat{t}_{\mathrm{Y}}^{cal}(t)) = \sum_{i \in \mathrm{U}} \sum_{j \in \mathrm{U}} (\pi_{ij} - \pi_i \pi_j) \frac{\mathrm{E}_i(t)}{\pi_i} \frac{\mathrm{E}_j(t)}{\pi_j}, \quad t \in [0, \mathrm{T}], \tag{2.26}$$

où $\mathrm{E}_i(t) = \mathrm{Y}_i(t) - \mathbf{X}_i' \boldsymbol{\beta}(t)$ est le résidu de la régression de $\mathrm{Y}_i(t)$ sur $\mathbf{X}_i$, avec

$$\boldsymbol{\beta}(t) = \left(\sum_{i \in \mathrm{U}} \mathbf{X}_i q_i \mathbf{X}_i' \right)^{-1} \sum_{i \in \mathrm{U}} \mathbf{X}_i q_i \mathrm{Y}_i(t), \quad t \in [0, \mathrm{T}].$$

Le calage sera donc d'autant plus efficace que les résidus pondérés $\frac{\mathrm{E}_i(t)}{\pi_i}$, $i \in \mathrm{U}$ seront petits en valeur absolue.

Un estimateur de cette variance est donné par :

$$\mathrm{A}\hat{\mathbb{V}}(\hat{t}_{\mathrm{Y}}^{cal}(t)) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{\hat{\mathrm{E}}_i(t)}{\pi_i}\frac{\hat{\mathrm{E}}_j(t)}{\pi_j}, \quad t \in [0,\mathrm{T}], \tag{2.27}$$

où $\mathrm{E}_i(t) = \mathrm{Y}_i(t) - \mathbf{X}_i'\hat{\boldsymbol{\beta}}(t)$.

De même, la fonction de covariance des estimateurs par calage aux instants t et u peut être approximée par

$$\mathrm{ACov}(\hat{t}_{\mathrm{Y}}^{cal}(t), \hat{t}_{\mathrm{Y}}^{cal}(u)) = \sum_{i\in \mathrm{U}}\sum_{i\in \mathrm{U}}(\pi_{ij}-\pi_i\pi_j)\frac{\mathrm{E}_i(t)}{\pi_i}\frac{\mathrm{E}_j(u)}{\pi_j}, \quad \forall t,u \in [0,\mathrm{T}]. \tag{2.28}$$

La convergence uniforme de la variance et la convergence ponctuelle de la covariance ont été étudiées dans Lardin (2012) et Cardot et al. (2013c).

Cas particulier du calage : estimateur de Hájek

Très souvent, on connaît la taille N de la population d'intérêt. Il peut donc être judicieux d'intégrer cette information dans l'estimation. Pour cela, on peut utiliser l'estimateur de Hájek qui peut être vu comme un estimateur par le calage qui utilise que la variable auxiliaire constante valant 1 pour chaque unité de la population.

L'estimateur de Hájek (voir Godambe and Sprott (1971)) est donné par

$$\hat{t}_{\mathrm{Y}}^{\mathrm{H}ajek}(t) = \frac{\mathrm{N}}{\sum_{j\in s}\frac{1}{\pi_j}}\frac{\mathrm{Y}_i}{\pi_i}, \quad t \in [0,\mathrm{T}]$$

et il donne en effet à chaque unité $i \in s$, le poids $w_{is} = \frac{\mathrm{N}}{\pi_i}\left(\sum_{j\in s}\frac{1}{\pi_j}\right)^{-1}$ tel que l'équation de calage sur cette variable constante est bien respectée :

$$t_{\mathbf{X}} = \sum_{i\in \mathrm{U}} 1 = \sum_{i\in s} w_{is}1 = \sum_{i\in s}\frac{\mathrm{N}}{\pi_i}\left(\sum_{j\in s}\frac{1}{\pi_j}\right)^{-1} = \mathrm{N}. \tag{2.29}$$

Cet estimateur a, en particulier, la propriété de corriger le défaut de l'estimateur d'Horvitz-Thompson de ne pas estimer de façon exacte le total d'une variable constante, i.e. $\mathrm{Y}_i = c$ pour tout $i \in \mathrm{U}$, pour un plan de sondages à probabilités inégales. En outre, Särndal (1992) (pages 182-184) donnent d'autres propriétés de l'estimateur de Hájek.

Dans le Chapitre 5 sur l'estimation en présence de valeurs manquantes, nous proposerons des extensions de cet estimateur ainsi que de l'estimateur de Horvitz-Thompson au cadre de l'estimation de courbes moyennes en présence de valeurs manquantes et nous commenterons les différences constatées entre ces estimateurs.

Approche basée sur le modèle

Jusqu'à présent, nous nous étions placés dans l'approche basée sur le plan de sondage et nous avions supposé que les valeurs des variables d'intérêt Y_i pour chaque unité de la population étaient déterministes et que le seul aléa présent était celui de la constitution de l'échantillon. L'inférence statistique décrivait alors uniquement le hasard engendré par le plan de sondage.

Cependant, dans le Chapitre 4, nous nous intéresserons à la problématique de l'estimation pour de petites sous-populations aussi appelées petits domaines. Or les méthodes usuellement utilisées pour répondre à ce type de problématique se basent fréquemment sur des modélisations du lien entre variables explicatives et variable d'intérêt construites à partir de l'ensemble de l'échantillon.

On va alors se placer dans ce qu'on appelle en sondage l'approche basée sur le modèle (ou approche *model-based*, décrite notamment dans Valliant et al. (2000)). Dans cette approche, on ne considère plus les quantités Y_i comme fixes mais comme aléatoires. L'aléa que l'on modélise alors n'est plus celui de la sélection de l'échantillon mais la la valeur des Y_i. On suppose donc l'existence de ce qu'on appelle un *modèle de superpopulation*, c'est-à-dire qu'on considère que les valeurs de la variable d'intérêt $(Y_1, \ldots, Y_i, \ldots, Y_N)$ pour les N unités de notre population U sont des réalisations indépendantes d'une même variable aléatoire $\mathscr{Y}$ de distribution ξ. On va donc chercher à modéliser cette distribution ξ, en particulier en décrivant le lien entre la variable d'intérêt et l'information auxiliaire. On note respectivement $\mathbb{E}_\xi$ et $\mathbb{V}_\xi$ l'espérance et la variance sous ce modèle.

Considérons un modèle de superpopulation non paramétrique

$$\xi : Y_i(t) = f(\mathbf{X}_i, t) + \epsilon_{it}, \quad t \in [0, T], \tag{2.30}$$

avec $f(\cdot)$ une fonction inconnue et les résidus ϵ_{it} sont des variables indépendantes par rapport à l'individu $i \in U$, de moyenne nulle, $\mathbb{E}_\xi(\epsilon_{it}) = 0$, et fonction de covariance $\mathrm{Cov}_\xi(\epsilon_{it}, \epsilon_{iu}) = \sigma_i(t, u)$ pour $t, u \in [0, T]$. Dans le contexte des petits domaines, nous verrons comment modifier ce modèle pour tenir compte des domaines auxquels appartiennent les unités.

Dans l'approche basée sur le modèle, la courbe totale t_Y à l'instant $t \in [0, T]$ peut être prédite à l'aide de l'estimateur proposé par Valliant et al. (2000), *i.e.*

$$\hat{t}_Y^{MB}(t) = \sum_{i \in s} Y_i(t) + \sum_{i \in U - s} \hat{f}(\mathbf{X}_i, t), \quad t \in [0, T],$$

où $\hat{f}$ est un estimateur de la fonction f. Tout l'enjeu est de trouver un estimateur $\hat{f}$ tel que, pour chaque $t \in [0, T]$, l'erreur quadratique moyenne (calculée sous le modèle) de l'erreur de prédiction $\hat{t}_Y^{MB}(t) - t_Y(t)$ soit la plus petite possible. De même, on définira l'estimateur de la moyenne

$$\hat{\mu}_Y^{MB}(t) = \frac{1}{N} \hat{t}_Y^{MB}(t), \quad t \in [0, T].$$

En particulier, dans le cas du modèle linéaire $Y_i(t) = \mathbf{X}_i' \boldsymbol{\beta}(t) + \epsilon_{it}, i \in U$, le meilleur (dans le sens donné ci-dessus) estimateur sans biais de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{i \in s} \frac{\mathbf{X}_i \mathbf{X}_i'}{\sigma_i(t, t)} \right)^{-1} \sum_{i \in s} \frac{\mathbf{X}_i Y_i(t)}{\sigma_i(t, t)}, \quad t \in [0, T]$$

et le meilleur estimateur sans biais $\hat{t}_{\mathrm{Y}}^{\mathrm{MB}}$ du total t_{Y} est donné par (voir Royall (1976)) :

$$\hat{t}_{\mathrm{Y}}^{\mathrm{MB}}(t) = \sum_{i \in s} \mathrm{Y}_i(t) + \sum_{i \in \mathrm{U}-s} \mathbf{X}_i' \hat{\boldsymbol{\beta}}(t) = \sum_{i \in s} w_{is} \mathrm{Y}_i(t),$$

où les poids sont donnés par

$$w_{is}(t) = 1 + \frac{\mathbf{X}_i'}{\sigma_i(t,t)} \left(\sum_{j \in s} \frac{\mathbf{X}_j \mathbf{X}_j'}{\sigma_j(t,t)} \right)^{-1} \left(\sum_{j \in \mathrm{U}-s} \mathbf{X}_j \right), \quad i \in s. \tag{2.31}$$

Cet estimateur est appelé aussi BLUP (Best Linear Unbiased Predictor). Lorsque l'on se place dans l'approche basée sur le modèle, un point de vigilance majeur concerne la qualité du modèle. En effet, comme les estimations reposent dans ce cadre sur la véracité de celui-ci, leur qualité dépend très fortement de la pertinence et de l'exactitude du modèle employé. Ainsi, l'utilisation de modèles erronés peut biaiser les estimateurs. Cela n'est pas le cas dans l'approche basée sur le plan dans laquelle l'inférence repose uniquement sur le réalisme des hypothèses faites sur le mécanisme de sélection de l'échantillon. Cependant, dans le contexte de l'estimation sur petits domaines, on espère que la diminution de variance engendrée par l'utilisation de modèles permettra des gains de précision qui surpasseront l'augmentation potentielle du biais en termes d'erreur quadratique moyenne.

Dans un cadre non-fonctionnel, Valliant et al. (2000) montrent que l'estimateur $\hat{t}_{\mathrm{Y}}^{\mathrm{MB}}$ est efficace quand $f(x)$ et $v(x)$ sont correctement spécifiés mais peut être biaisé si le modèle est faux. Cependant, Valliant et al. (2000) proposent des méthodes pour rendre plus robustes les estimateurs à une mauvaise spécification du modèle,

Toujours dans un cadre non-fonctionnel et pour un estimateur non paramétrique à noyau de la fonction de régression, Dorfman (1994) donne des conditions sous lesquelles le biais relatif de l'estimateur basé sur un modèle ainsi que la variance tendent vers 0. En outre, sous des conditions supplémentaires, ils montrent que l'erreur de prédiction $\hat{t}_{\mathrm{Y}}^{\mathrm{MB}} - t_{\mathrm{Y}}$ est asymptotiquement normale.

Dans le contexte EDF, la sélection d'échantillons est fréquemment soumise à des contraintes techniques fortes. Par exemple, pour des raisons de coût, nous sélectionnons les unités échantillonnées parmi les clients équipés de compteurs communicants Linky. Or le déploiement des compteurs communicants, dont le but premier n'est pas la collecte de données à visée statistique mais l'amélioration de la gestion du réseau de distribution d'électricité, est actuellement en cours et s'effectue par grappe, quartier par quartier. Durant cette période transitoire de déploiement, il peut donc exister des sur-représentations ou sous-représentations de certains types de communes (urbaines notamment) ou encore de certaines zones climatiques. Ces caractéristiques pouvant avoir un impact sur les consommations électriques, il faudra prendre soin de les inclure dans nos modélisations.

2.2.5 Estimation de la variance sous le plan

Dans ce paragraphe, on s'intéresse aux estimateurs ponctuels de variance, c'est-à-dire à des estimateurs de la variance d'un estimateur de la courbe moyenne ou totale pour un instant t donné. La construction de bandes de confiance n'est pas abordée ici

mais a été traitée notamment par Cardot et al. (2013a), Cardot et al. (2013b) et Cardot et al. (2013c).

Pour des estimateurs de totaux ou de moyennes et pour des plans de sondage dont les probabilités d'inclusion simples et doubles sont connues (par exemple le sondage aléatoire stratifié), des estimateurs ou des approximations de variance analytiques peuvent être fournis (voir les expressions (2.9) et (2.26)). Cependant, lorsque l'on s'intéresse à des estimateurs plus complexes (par exemple les estimateurs robustes avec projection sur la base de l'Analyse en Composantes Principales Sphériques présentés dans le Chapitre 3) ou encore lorsque l'on réalise des estimations en présence de non réponse comme dans le Chapitre 5, il ne sera pas toujours possible de trouver un estimateur analytique de la variance. On aura donc recours à des techniques d'approximations, et notamment la linéarisation et le bootstrap que nous présentons ci-dessous.

Linéarisation en sondages

La linéarisation permet d'approximer la précision de l'estimateur d'une statistique non linéaire (par exemple un ratio) de totaux en se ramenant à un problème d'estimation de la variance pour un total ou une moyenne. L'idée générale est d'appliquer un développement de Taylor de l'estimateur afin d'en obtenir une approximation. On estime ensuite la variance de cette approximation. Dans le cadre de de la théorie des sondages, Särndal et al. (2003) a détaillé la linéarisation comme un développement de Taylor alors que Deville (1999) a proposé une approche via la fonction d'influence et les variables linéarisées.

Plus précisément, soit $\theta = f(t_{Y_1}, \ldots, t_{Y_\Omega})$ le paramètre d'intérêt, fonction non-linéaire de totaux, et $\hat{\theta} = f(\hat{t}_{Y_1}, \ldots, \hat{t}_{Y_\Omega})$ son estimateur par substitution. On suppose que f est dérivable et homogène de degré $\alpha \geq 0$, c'est-à-dire $f(cx) = c^\alpha f(x)$ pour tout x réel. On suppose également que $\lim_{N\to\infty} N^{-\alpha}\theta < \infty$ et que $N^{-1}(\hat{t}_{Y_\omega} - t_{Y_\omega}) = O_p\left(\frac{1}{\sqrt{n}}\right)$ pour tout $\omega = 1, \ldots, \Omega$. En utilisant un développement de Taylor, on obtient

$$\begin{aligned} N^{-\alpha}\hat{\theta} &= N^{-\alpha}\theta + N^{-\alpha} \sum_{\omega=1}^{\Omega} \frac{\partial f(x_1, \ldots, x_\Omega)}{\partial x_\omega}\Big|_{\left(x_\ell = t_{Y_\ell}\right)_{\ell=1}^{\Omega}} (\hat{t}_{Y_\omega} - t_{Y_\omega}) + o(||N^{-1}(\hat{\theta} - \theta)||) \\ &= N^{-\alpha}\theta + N^{-\alpha}\left(\sum_{i\in s} \frac{u_i}{\pi_i} - \sum_{i\in U} u_i\right) + o_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \tag{2.32}$$

où

$$u_i = \sum_{\omega=1}^{\Omega} \frac{\partial f(x_1, \ldots, x_\Omega)}{\partial x_\omega}\Big|_{\left(x_\ell = t_{Y_\ell}\right)_{\ell=1}^{\Omega}} Y_{i\omega}$$

est la variable linéarisée de θ calculée pour l'individu $i \in U$.

On peut ensuite en déduire une approximation la variance de $\hat{\theta}$

$$\mathbb{V}_{approx}(\hat{\theta}) = \sum_{i\in U} \sum_{j\in U} (\pi_{ij} - \pi_i \pi_j) \frac{u_i}{\pi_i} \frac{u_j}{\pi_j}. \tag{2.33}$$

En pratique, les valeurs u_i ne sont pas connues même pour les individus échantillonnés puisqu'elles dépendent des valeurs des variables d'intérêt Y_i et on les estime

par $\hat{u}_i = \sum_{\omega=1}^{\Omega} \frac{\partial f(x_1,\ldots,x_\Omega)}{\partial x_\omega}\Big|_{\left(x_\ell=\hat{t}_{Y_\ell}\right)_{\ell=1}^{\Omega}} Y_{i\omega}$. Un estimateur de cette variance est alors donné par

$$\hat{\mathbb{V}}_{approx}(\hat{\theta}) = \sum_{i\in s}\sum_{j\in s} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \frac{\hat{u}_i}{\pi_i}\frac{\hat{u}_j}{\pi_j}. \tag{2.34}$$

Bootstrap en sondages

Le bootstrap est une méthode de rééchantillonnage permettant d'approximer la variance de statistiques complexes. Il a été introduit par Efron (1979) pour une population infinie. L'adaptation de cette technique au cadre de la population finie a fait l'objet d'une large littérature. En effet, il n'existe pas de méthode générale de bootstrap pour un plan de sondage quelconque mais plutôt des propositions pour chaque type de plan. Pour un tour d'horizon de ce sujet riche, on pourra se référer notamment à Shao and Tu (2012), Chauvet (2007).

Le bootstrap consiste à estimer la distribution de l'estimateur $\hat{\theta}$ de notre paramètre d'intérêt θ, inconnue, par sa distribution empirique constatée sur les données. Pour cela, l'algorithme du bootstrap consiste à générer un grand nombre de "rééchantillons" s^* à partir de l'échantillon de départ s, à calculer l'estimateur $\hat{\theta}^*$ sur chaque rééchantillon puis à considérer la distribution des estimateurs obtenus sur chacun des rééchantillons comme une approximation de la distribution de l'estimateur. En particulier, la variance de l'estimateur pourra être approximée par la variabilité des estimateurs bootstrap entre les différents rééchantillons. Ainsi, soit B le nombre de rééchantillons, un estimateur de variance par bootstrap est

$$\hat{\mathbb{V}}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*(b)} - \overline{\hat{\theta}^*})^2, \tag{2.35}$$

où $\hat{\theta}^{*(b)}$ est l'estimateur obtenu sur le rééchantillon b et $\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)}$ la moyenne de ces estimateurs sur l'ensemble des B réplications.

La question qui se pose alors dans le contexte de l'estimation par sondage en population finie est de savoir comment générer les rééchantillons. Différents algorithmes ont été proposés pour répondre à cette problématique, notamment le *Rescaling Bootstrap* de Rao and Wu (1988), le *Mirror-Match Bootstrap* de Sitter (1992), le bootstrap généralisé proposé par Bertail and Combris (1997) ou encore le *Bootstrap populationnel* proposé par Gross (1980). Ce dernier a ensuite été amélioré par Booth et al. (1994) pour tenir compte du fait que les inverses des probabilités d'inclusion ne sont pas forcément entières. Dans le Chapitre 3, nous suivrons les approches du bootstrap généralisé et du bootstrap populationnel pour approximer la variance de nos estimateurs robustes, c'est pourquoi nous allons présenter ces deux méthodes dans la suite de ce paragraphe.

Le bootstrap populationnel Le principe du bootstrap populationnel est de générer une pseudo population U^* selon le principe de Horvitz-Thompson en répliquant chaque unité i de l'échantillon un nombre de fois égal à son poids de sondage puis ensuite à tirer les rééchantillons s^* dans cette pseudo population selon le même plan

de sondage que celui ayant conduit à la constitution de s. Dans le cas du sondage aléatoire simple, pour tenir compte du fait que les poids de sondage ne sont pas forcément entiers, Booth et al. (1994) proposent de répliquer chaque unité un nombre de fois égal à la partie entière de son poids de sondage puis de compléter la pseudo population par tirage aléatoire simple pour arriver à la taille N de la population de départ. La pseudo population devient alors également aléatoire et doit donc être simulée à nouveau pour chaque rééchantillon. Cette idée est étendue aux plans de sondages à probabilités inégales dans Chauvet (2007) et Holmberg (1998).

Le bootstrap généralisé Le bootstrap généralisé proposé par Bertail and Combris (1997) et étudié notamment par Beaumont and Patak (2012) et Antal and Tillé (2011), consiste à répliquer non pas les valeurs Y_i dans l'échantillon mais les poids des individus dans chaque rééchantillon, en générant ces poids de façon à respecter des contraintes sur leurs moments permettant de reproduire l'aléa induit par plan de sondage. L'objectif du bootstrap est d'approximer la mesure empirique.

Plus précisément, pour chaque rééchantillon $b = 1, \ldots, \mathrm{B}$, on génère aléatoirement l'ensemble des n poids de réplications $\mathrm{W}_i^{*(b)}, \quad i \in s$, tels que (voir Bertail and Combris (1997), Section 2.2)

$$\begin{aligned}
\mathbb{E}(\mathrm{W}_i^{*(b)}) &= \frac{1}{\mathrm{N}}, \quad \forall i \in s, \\
\mathbb{V}(\mathrm{W}_i^{*(b)}) &= \frac{1}{\mathrm{N}^2}(1-\pi_i), \quad \forall i \in s, \\
Cov(\mathrm{W}_i^{*(b)}, \mathrm{W}_j^{*(b)}) &= \frac{1}{\mathrm{N}^2}(1-\frac{\pi_i \pi_j}{\pi_{ij}}) \quad \forall i, j \in s, \quad i \neq j.
\end{aligned}$$

On peut par exemple utiliser une loi normale multivariée. C'est ce que nous ferons dans le Chapitre 2.

Ensuite, pour chaque réplication bootstrap, on obtient l'estimateur de la moyenne par

$$\hat{\mu}_{\mathrm{Y}}^{*(b)} = \sum_{i \in s} \mathrm{W}_i^{*(b)} d_i \mathrm{Y}_i,$$

où $d_i = \frac{1}{\pi_i}, i \in s$ est le poids de Horvitz-Thompson de l'unité i. De même, l'estimateur du total est obtenu par

$$\hat{t}_{\mathrm{Y}}^{*(b)} = \mathrm{N} \sum_{i \in s} \mathrm{W}_i^{*(b)} d_i \mathrm{Y}_i.$$

On peut ensuite en déduire l'estimateur de la variance par l'équation (2.35).

Chapitre 3

Estimation par sondage de courbes moyennes ou totales de consommation électrique robuste aux unités influentes

3.1 Contexte et introduction

Que l'on s'intéresse à des clients résidentiels, professionnels ou industriels, les consommations d'électricité ont en général une distribution très asymétrique, avec beaucoup de très petits clients et quelques très gros. Les estimateurs classiques de courbes de consommation moyennes ou totales peuvent être très sensibles à la présence de ces "gros clients" dans l'échantillon. Cet effet est d'autant plus marqué lorsque nous nous intéressons à de petites sous-populations pour lesquelles nous disposons de peu d'individus dans les panels : en effet plus l'échantillon utilisé est petit, plus une unité isolée peut avoir une influence importante à elle seule sur l'estimation. Pour tenter de réduire l'impact d'éventuels individus atypiques sur la qualité de nos estimations, nous proposons donc dans ce chapitre des estimateurs de courbes de charge moyennes ou totales robustes à ce qu'on appelle en sondages des *unités influentes*, c'est-à-dire des unités dont la présence ou l'absence dans l'échantillon a un impact substantiel sur les estimateurs. (voir Chambers (1986) pour une définition plus précise). On s'intéresse ici uniquement à ce que Chambers (1986) appelle les *representative outliers*, c'est-à-dire les unités atypiques résultant d'une valeur inhabituelle mais exacte de la variable d'intérêt et non pas d'une erreur dans le processus de collecte. La question de l'exactitude des données relevées et du processus de détection et de correction d'éventuelles erreurs ne sera pas traitée ici.

De manière générale, une unité est considérée comme influente si, pour une configuration donnée (population d'étude et variable d'intérêt, paramètre, estimateur et plan de sondage), sa valeur a un impact important sur la variance de l'estimateur (Favre-Martinoz (2015)).

Lorsque l'on travaille sur des données de consommation électrique, l'apparition d'outliers, c'est-à-dire d'unités atypiques, est très probable du fait de la forte asymétrie de la distribution des consommations observées dans la population mais aussi de la nature très chahutée des courbes de charge, qui peut aboutir à ce qu'on appelle des *peak outliers* (i.e. des unités dont le niveau moyen est normal mais dont la courbe at-

teint des valeurs anormales pour certains intervalles de temps).

Dans le contexte de l'estimation en population finie, une bonne façon de limiter l'impact de ces outliers est d'utiliser un plan de sondage stratifié (voir 2.2.4). Cependant, les panels peuvent contenir ce qu'on appelle des *strata jumpers* c'est-à-dire des unités très différentes des autres unités de leur strate du fait d'une base de sondage inexacte (par exemple, le code NAF des entreprises, qui décrit leur secteur d'activité, peut être erroné) ou encore de changements récents dans les caractéristiques des unités échantillonnées (par exemple un changement de composition familiale ou un changement de mode de chauffage pour des clients particuliers). La présence de ces *strata jumpers* détériore fortement l'homogénéité de celle-ci et donc la précision de l'estimateur associé.

Une unité peut également être influente car elle a un grand poids, par exemple suite à un calage si l'échantillon présentait des biais de sélection ou de non réponse forts qu'il a fallu redresser. Enfin, il peut également arriver qu'un individu atypique ait également un grand poids, ce qui rend son influence sur l'estimateur de moyenne ou de total d'autant plus grande.

On voit donc que, dans le cadre de l'estimation par sondage en population finie, la notion d'influence a un sens différent de celui donné en statistique classique. En effet, en population finie, même une unité non atypique peut être influente, par exemple si elle a un poids important du fait du plan de sondage. On fait donc la distinction entre un outlier (unité atypique) et une unité influente. En outre, en population finie, une unité non échantillonnée peut également être influente.

L'impact des unités influentes pourra être d'autant plus important que la taille de l'échantillon considéré est petite : en effet, une unité isolée pourra alors à elle seule avoir d'autant plus d'influence sur l'estimation. Cela est le cas notamment lorsqu'on souhaite estimer la courbe moyenne d'une petite sous-population (par exemple l'ensemble des individus possédant un équipement électrique particulier relativement rare, tel qu'une pompe à chaleur) : nos panels ayant été dimensionnés pour assurer une qualité satisfaisante pour les études concernant l'ensemble de la population, on ne dispose alors que de peu d'individus appartenant à cette sous-population particulière. Dans certains cas, si l'on s'intéresse à *plusieurs sous-populations* simultanément et si ces sous-populations sont suffisamment homogènes pour que l'on puisse poser un modèle commun sur l'ensemble de celles-ci, on pourra alors utiliser des modèles de type "petits domaines", présentés dans le chapitre suivant 4 pour améliorer la qualité des estimations. En particulier, ces méthodes pourront s'appliquer lorsque l'on souhaite disposer d'estimations de courbes moyennes à la maille de chaque zone géographique.

Il existe cependant des cas où cela ne sera pas possible et on devra alors travailler sur de petits sous-échantillons. Par exemple, si on travaille sur des clients professionnels, il serait insensé de poser un modèle commun sur les boulangeries, les discothèques et les aciéries. De même, si l'on s'intéresse aux clients disposant de pompes à chaleur, cela ne serait pas pertinent d'utiliser dans les estimations des clients équipés d'autres moyens de chauffage.

Dans le cadre de l'estimation en population finie, le traitement des unités influentes est d'autant plus délicat que, contrairement au cadre des travaux en population finie, on ne s'intéresse pas uniquement au comportement des unités "non aty-

piques" mais bien à l'estimation de quantités totales ou moyennes sur l'ensemble de la population, y compris les outliers. Ceux-ci ne sont donc pas uniquement des "nuisances à éliminer". Hors du contexte des données fonctionnelles, il existe des estimateurs robustes aux unités influentes en sondages. On pourra citer par exemple Chambers (1986), Gwet and Rivest (1992), Rivest (1994), Kokic and Bell (1994), Welsh and Ronchetti (1998). De manière très schématique, on peut considérer que ces estimateurs sont basés sur des techniques de winsorisation [1] permettant de limiter les influences des unités selon une fonction dépendant d'une constante d'ajustement dont le choix est crucial. On pourra se référer au chapitre 11 de Pfeffermann and Rao (2009) pour une présentation plus détaillée des méthodes d'estimation en présence d'outliers ou de valeurs influentes en sondages.

Récemment, Beaumont et al. (2013) ont proposé une approche unifiée permettant de transformer des estimateurs non robustes en estimateurs robustes en bornant l'influence de chaque unité échantillonnée sur l'estimateur, cette influence étant mesurée par une quantité appelée biais conditionnel et la limite de troncature étant déterminée selon une approche minimax. Le biais conditionnel a été introduit par Muñoz-Pichardo et al. (1995) dans le contexte des modèles linéaires généraux et repris par Moreno-Rebollo et al. (1999) dans le contexte des estimations par sondage. Plus précisément, pour transformer un estimateur non robuste en estimateur robuste, il s'agit de faire apparaître les expressions des biais conditionnels de chacune des unités échantillonnées dans l'expression de cet estimateur, puis de tronquer ces biais conditionnels, c'est-à-dire de les borner de façon à ce qu'ils n'excèdent pas un certain seuil. L'influence de chaque unité sur l'estimateur est alors limitée par construction, ce qui confère à ce dernier la robustesse souhaitée. La principale question est alors de trouver le seuil optimal de troncature et les auteurs y répondent à l'aide d'une approche minimax.

L'objectif de ce chapitre est de construire des estimateurs de courbes moyennes par sondage, robustes à la présence d'unités influentes dans les échantillons. Pour cela, nous proposons d'adapter de différentes manières l'approche proposée par Beaumont et al. (2013) au cadre des données fonctionnelles. En effet, dans ce cadre l'estimateur ainsi que les biais conditionnels sont des courbes. La question centrale de ce chapitre sera donc de savoir comment "tronquer des courbes", ou plus précisément rechercher la fonction de troncature fonctionnelle à appliquer sur nos biais conditionnels, afin de tirer parti du mieux possible du caractère fonctionnel de notre problématique. De manière plus générale, on cherche la meilleure façon possible de découper un problème de dimension infinie (dans ce chapitre la mesure d'influence et la recherche du meilleur estimateur robuste) en un nombre fini de problèmes "parallèles" plus simples et déjà traités dans la littérature que l'on pourra résoudre de manière indépendante sans casser la cohérence intrinsèque de notre problématique.

Après avoir introduit quelques notions sur la robustesse en statistique et les mesures d'influence dans la section 3.2, nous nous intéressons dans la section 3.3 aux biais conditionnels, qui constituent notre mesure d'influence et sont construits en déclinant l'approche de Beaumont et al. (2013) dans le cadre fonctionnel. On présente ensuite dans 3.4 trois méthodes robustes d'estimation de courbes totales par sondage.

1. seuillage des valeurs extrêmes

La première, présentée au paragraphe 3.4.1 consiste à employer l'estimateur robuste de totaux réels proposé par Beaumont et al. (2013) indépendamment sur les valeurs des courbes pour chacun des instants de discrétisation. Cependant, on se doute que cette approche n'est pas optimale car elle n'exploite pas les fortes corrélations temporelles de nos courbes. Nous proposons donc également deux autres estimateurs qui prennent pleinement en compte l'aspect fonctionnel du problème. Le premier de ces estimateurs, décrit dans le paragraphe 3.4.2, consiste à projeter les courbes sur une base de dimension finie pertinente (par exemple sur les composantes principales obtenues en appliquant une variante robuste de l'ACP, appelée Analyse en Composantes Principales Sphériques) sur les courbes discrétisées ou encore une base d'ondelettes. La seconde solution, décrite dans le paragraphe 3.4.3, consiste à tronquer de manière globale les biais conditionnels, c'est-à-dire à trouver les limites de troncature simultanément sur l'ensemble des instants de la période d'étude. Pour cela, on se base sur la notion de profondeur fonctionnelle qui est une mesure du caractère atypique d'une courbe. Dans le paragraphe 3.4.4, on propose par ailleurs un critère alternatif à celui du minimax utilisé par Beaumont et al. (2013) basé sur la minimisation d'une fonction différente des biais conditionnels. Il peut se décliner aussi bien dans le contexte univarié que pour nos deux approches fonctionnelles. Ensuite, des estimateurs approximant l'erreur quadratique moyenne à chaque instant de nos différents estimateurs robustes sont proposés dans la section 3.5 soit en se basant sur des bootstraps populationnels ou généralisés, soit à l'aide d'une expression explicite obtenue en considérant comme fixes les constantes de troncature. Enfin, l'ensemble de ces estimateurs robustes ainsi que les estimateurs d'erreur quadratique moyenne associés sont comparés entre eux sur des jeux de données réels dans la section 3.6.

3.2 Robustesse en statistique

Dans cette section, on présente quelques éléments de bibliographie et concepts relatifs à la robustesse en statistique, dans un contexte général c'est-à-dire non spécifique à l'estimation par sondage en population finie. Ces différents éléments, même s'ils ne sont pas directement liés à notre problématique, peuvent faciliter la compréhension des intuitions qui sous-tendent notre démarche. Les notions présentées ici sont tirées principalement de Hampel et al. (2011).

L'objectif de la statistique robuste est de produire des estimateurs dont les performances sont proches de celles des estimateurs non robustes lorsque le modèle est vérifié, mais qui en outre ont de bonnes performances pour de "petites déviations" par rapport à ce modèle. Ces petites déviations peuvent signifier que le modèle considéré est un peu différent de ce qui a été spécifié ou encore que l'hypothèse d'indépendance des données est remise en cause, ou bien qu'une petite partie de l'échantillon (les outliers) ne suit pas la loi donnée. Dans notre contexte des sondages, c'est ce dernier cas qui nous intéresse plus particulièrement.

L'ouvrage de Huber (1964) constitue une des références sur la statistique robuste. La notion de *robustesse qualitative*, qui y est définie signifie qu'une petite déviation par rapport au modèle (dans notre cas un petit nombre d'unités influentes) doit engendrer une petite modification de l'estimateur. Un estimateur robuste doit donc être bon "aux

environs du modèle" (ici pour un petit nombre d'unités influentes) et "pas tellement plus mauvais" qu'un estimateur non robuste lorsque le modèle est vérifié.

On risque donc de perdre en optimalité lorsque le modèle est exact, mais de gagner dans tous les autres cas. Ce compromis s'illustre en général par un compromis biais/variance : les estimateurs robustes sont fréquemment de plus faible variance mais au prix d'un biais potentiel. L'enjeu sera donc de trouver le paramétrage optimal de la méthode utilisée afin de minimiser un critère qui combine biais et variance, en l'occurrence l'erreur quadratique moyenne.

Pour qu'un estimateur soit robuste, une des exigences est qu'une unité ne puisse pas avoir à elle seule un trop grand impact sur l'estimateur final. Ainsi un nombre réduit d'unités influentes ne doit pas suffire à rendre trop instables les estimations. En population infinie, cette notion d'impact d'une unité peut être mesurée à l'aide de la *fonction d'influence*.

Ainsi, considérons une variable d'intérêt Y de distribution F. On dispose d'un échantillon $Y_1, \ldots, Y_n$ de n observations indépendantes identiquement distribuées. On s'intéresse à une fonction de la variable d'intérêt $\theta = \theta(F)$. La fonction d'influence représente l'influence incrémentale d'un nouveau point en Y_i sur $\theta(F)$:

$$\mathrm{IF}(\theta, F, Y_i) = \frac{1}{\iota} \lim_{\iota \longrightarrow 0} \left(\theta\Big((1-\iota)F + \iota\delta_{Y_i}\Big) - \theta(F) \right). \tag{3.1}$$

D'après notre définition de la robustesse, il faut donc que cette fonction d'influence soit bornée. Dans la suite, nous verrons que nous utiliserons plutôt une autre mesure d'influence, appelée biais conditionnel, plus adaptée au contexte de l'estimation en population finie. En effet, en population infinie, l'objectif est généralement d'estimer un paramètre de la population des inliers alors qu'en sondages on s'intéresse à des paramètres de l'ensemble de la population y compris les outliers. Des critères de robustesse comme une fonction d'influence bornée ou un point de rupture élevé ne sont donc pas forcément les plus pertinents.

Parmi les estimateurs robustes, on peut citer les **M-estimateurs**, proposés par Hampel et al. (2011), qui sont une extension des méthodes de maximisation telles que le maximum de vraisemblance. Ces méthodes consistent à trouver l'estimateur de θ qui minimise $\sum_{i=1}^{n} \rho(\mathbf{X}_i, \theta^*)$. Cela revient à trouver θ^* tel que

$$\sum_{i=1}^{n} \Psi(\mathbf{X}_i, \theta^*) = 0$$

avec Ψ la dérivée de ρ par rapport à θ.

Pour ces estimateurs, la fonction d'influence est proportionnelle à Ψ. Pour qu'elle soit bornée et donc que l'estimateur soit robuste, il suffit donc que Ψ soit bornée.

Les estimateurs de Huber proposés dans Huber (1964) pour l'estimation du paramètre de tendance centrale d'une variable Y réelle constituent un cas particulier de M-estimateurs. La fonction Ψ utilisée est alors de la forme :

$$\psi_c(x) = \begin{cases} x & \text{si} \quad x \in [-c; c] \\ c & \text{si} \quad x > c \\ -c & \text{si} \quad x < -c, \end{cases} \tag{3.2}$$

avec $c \in \mathbb{R}^+$ une constante qu'il faut choisir judicieusement : plus c augmente, plus on se rapproche de la moyenne (dont on peut montrer aisément qu'elle correspond au choix de la fonction identité $\psi(x) = x$) : l'influence n'est alors pas bornée, l'estimateur n'est pas robuste, mais il n'est pas biaisé. A l'inverse, lorsque c tend vers 0, l'estimateur tend vers la médiane (voir Hampel et al. (2011)). Le choix de c permet donc un équilibre entre robustesse et biais. Cet estimateur est également celui obtenu en suivant l'**approche minimax** proposée par Huber (1964). Huber présente cet estimateur minimax comme la solution optimale d'un jeu dans lequel "la nature" choisit la pire distribution possible au sens de l'information de Fisher dans un voisinage donné du modèle, et le statisticien choisit le meilleur estimateur (i.e. celui de plus faible variance) pour cette pire distribution. Nous verrons que cette notion de minimax est utilisée pour le choix de la constante de troncature dans les méthodes robustes en sondages.

Enfin, rappelons qu'il existe des différences notables entre la robustesse en statistique classique, où l'on cherche à limiter l'impact des outliers et la robustesse en population finie, où l'on s'intéresse également aux unités influentes.

Dans le cadre de la statistique en population infinie, Muñoz-Pichardo et al. (1995) définissent le biais conditionnel comme une mesure de l'influence d'une unité i sur un estimateur θ, de la forme :

$$\mathrm{B}_i(y_i;\theta) = \mathrm{E}(\hat{\theta}|\mathrm{Y}_i = y_i) - \theta. \tag{3.3}$$

3.3 Biais conditionnel pour des courbes

Dans cette section, nous présentons les expressions des biais conditionnels ainsi que de leurs estimateurs pour différents estimateurs de courbes totales.

On se place dans le cadre de travail défini dans le paragraphe 2.2.1. Dans la suite, le paramètre d'intérêt considéré est le total t_Y sur une population U de la variable Y, mais la démarche s'étend très aisément au cas où on s'intéresse plutôt à la moyenne μ_Y. On se place dans l'approche basée sur le plan de sondage. On considère en particulier le plan de sondage aléatoire simple sans remise et le plan de sondage stratifié présentés dans 2.2.4. On s'intéresse tout spécialement à l'estimateur de Horvitz-Thompson (2.8) et l'estimateur par calage (2.23).

Dans l'approche basée sur le plan, le biais conditionnel d'une unité i pour un estimateur $\hat{\theta}$ représente alors l'espérance de cet estimateur conditionnellement à l'indicatrice d'inclusion I_i de l'unité i. En effet, pour une unité échantillonnée i, on a $\mathrm{B}_{1i}^{\hat{\theta}} = \mathrm{E}_p(\hat{\theta}|\mathrm{I}_i = 1) - \theta$ et pour une unité non échantillonnée on a $\mathrm{B}_{0i}^{\hat{\theta}} = \mathrm{E}_p(\hat{\theta}|\mathrm{I}_i = 0) - \theta$.

Dans notre contexte où la quantité d'intérêt est le total t_Y d'une variable Y sur une population, le biais conditionnel s'écrit, pour une unité échantillonnée.

$$\mathrm{B}_{1i}^{\hat{t}_\mathrm{Y}}(t) = \mathrm{E}_p(\hat{t}_\mathrm{Y}(t)|\mathrm{I}_i = 1) - t_\mathrm{Y}(t), \quad t \in [0,\mathrm{T}], \quad i \in s \tag{3.4}$$

Pour une unité non échantillonnée, il est égal à :

$$\mathrm{B}_{0i}^{\hat{t}_\mathrm{Y}}(t) = \mathrm{E}_p(\hat{t}_\mathrm{Y}(t)|\mathrm{I}_i = 0) - t_\mathrm{Y}(t), \quad t \in [0,\mathrm{T}], \quad i \in \mathrm{U} - s \tag{3.5}$$

Pour certains plans de sondage, l'erreur d'échantillonnage sur le total est approximativement égale à la somme des biais conditionnels. Celui-ci peut donc être inter-

prété comme la contribution de cette unité à l'erreur d'échantillonnage (voir Beaumont et al. (2013)). Par conséquent, le fait de borner l'influence de chaque unité rendra l'estimateur plus stable, car alors une unité isolée ne pourra avoir qu'un effet limité sur l'estimation.

Par exemple, pour un sondage stratifié et $\hat{t}_{\mathrm{Y}}^{\mathrm{HT}}$ l'estimateur de Horvitz-Thompson du total d'une courbe Y, le biais conditionnel d'une unité échantillonnée i appartenant à la strate h est (voir Beaumont et al. (2013)) :

$$\mathrm{B}_{1i}^{\hat{t}_{\mathrm{Y}}^{\mathrm{HT}}}(t) = \frac{\mathrm{N}_h}{\mathrm{N}_h - 1}\left(\frac{\mathrm{N}_h}{n_h} - 1\right)(\mathrm{Y}_i(t) - \overline{\mathrm{Y}}_h(t)), \quad t \in [0, \mathrm{T}], \tag{3.6}$$

où $\overline{\mathrm{Y}}_h$ est la moyenne de Y pour la strate U_h de la population. Cette expression montre bien que le biais conditionnel d'une unité et donc son influence dépendent de son caractère atypique en comparaison des autres unités de sa strate (ici la différence entre la valeur Y_i et la moyenne de Y dans la strate) ainsi que de son poids $\frac{\mathrm{N}_h}{n_h}$: plus le poids de l'unité sera important et plus l'unité sera différente des autres unités de sa strate, plus son influence sera grande. Le biais conditionnel dépend de paramètres inconnus de la population (en l'occurrence les moyennes des strates $\overline{\mathrm{Y}}_h$) qui doivent être estimés. L'estimateur de biais conditionnel qui en découle devient alors

$$\hat{\mathrm{B}}_{1i}^{\hat{t}_{\mathrm{Y}}^{\mathrm{HT}}}(t) = \frac{\mathrm{N}_h}{\mathrm{N}_h - 1}\left(\frac{\mathrm{N}_h}{n_h} - 1\right)(\mathrm{Y}_i(t) - \hat{\overline{\mathrm{Y}}}_h(t)), \quad t \in [0, \mathrm{T}], \tag{3.7}$$

avec $\hat{\overline{\mathrm{Y}}}_h$ la moyenne de Y dans l'échantillon s_h.

Pour un estimateur par calage, le biais conditionnel d'une unité échantillonnée devient

$$\mathrm{B}_{1i}^{\hat{t}_{\mathrm{Y}}^{cal}}(t) \simeq \frac{\mathrm{N}_h}{\mathrm{N}_h - 1}\left(\frac{\mathrm{N}_h}{n_h} - 1\right)\left(\mathrm{E}_i(t) - \overline{\mathrm{E}}_h(t)\right), \quad t \in [0, \mathrm{T}], \tag{3.8}$$

avec E_i les résidus de la régression linéaire fonctionnelle (voir 2.1.4) des Y_i sur les variables de calage $\mathbf{X}_i$.

Il peut être estimé par

$$\hat{\mathrm{B}}_{1i}^{\hat{t}_{\mathrm{Y}}^{cal}}(t) \simeq \frac{n_h}{n_h - 1}\left(\frac{\mathrm{N}_h}{n_h} - 1\right)\hat{\mathrm{E}}_i(t), \quad t \in [0, \mathrm{T}], \tag{3.9}$$

avec $\hat{\mathrm{E}}_i(t) = \mathrm{Y}_i(t) - \hat{\boldsymbol{\beta}}'(t)\mathbf{X}_i$ et $\hat{\boldsymbol{\beta}}(t) = \left(\sum_{i \in s} \frac{\mathbf{X}_i \mathbf{X}_i'}{\pi_i}\right)^{-1} \sum_{i \in s} \frac{\mathbf{X}_i \mathrm{Y}_i(t)}{\pi_i}$, l'estimateur des moindres carrés ordinaires, et $q_i = 1 \forall i \in \mathrm{U}$ [2]

3.4 Estimation robuste de la courbe totale

Dans cette section, nous allons voir comment utiliser les biais conditionnels définis dans la section précédente afin de rendre nos différents estimateurs de courbes totales robustes aux unités influentes. Notre but est d'adapter l'approche unifiée proposée

2. On peut également remplacer l'estimateur des moindres carrés par un estimateur par régression robuste (voir Huber (1964)) afin d'être moins sensible à la présence d'unités influentes.

par Beaumont et al. (2013) au cas où la variable d'intérêt Y, et donc les estimateurs et les biais conditionnels, sont fonctionnels. On cherchera en particulier à construire des estimateurs robustes qui prennent en compte l'aspect fonctionnel de notre problématique et préservent la cohérence temporelle de la courbe moyenne ou totale estimée.

Les biais conditionnels, dont nous venons de présenter quelques exemples, peuvent nous permettre de construire des estimateurs robustes $\hat{\theta}^{R}(t)$ à partir d'estimateurs non robustes $\hat{\theta}(t)$. En s'inspirant de Beaumont et al. (2013), on peut montrer que pour l'estimateur de Horvitz-Thompson et pour des plans de sondage stratifiés ou de taille fixe à forte entropie avec des probabilités d'inclusion inégales, nous avons

$$\hat{\theta}(t) - \theta(t) \approx \sum_{i \in s} B_{1i}^{\hat{\theta}}(t) + \sum_{i \in U-s} B_{0i}^{\hat{\theta}}(t). \tag{3.10}$$

Cette approximation devient une égalité pour le plan de Poisson et un total.

Ecrivons maintenant l'estimateur non-robuste comme suit :

$$\hat{\theta}(t) = \left(\hat{\theta}(t) - \sum_{i \in s} B_{1i}^{\hat{\theta}}(t) - \sum_{i \in U-s} B_{0i}^{\hat{\theta}}(t)\right) + \sum_{i \in s} B_{1i}^{\hat{\theta}}(t) + \sum_{i \in U-s} B_{0i}^{\hat{\theta}}(t) \tag{3.11}$$

et en utilisant (3.10), on obtient

$$\hat{\theta}(t) \simeq \theta(t) + \sum_{i \in s} B_{1i}^{\hat{\theta}}(t) + \sum_{i \in U-s} B_{0i}^{\hat{\theta}}(t). \tag{3.12}$$

Donc, le premier terme de l'expression (3.12) n'est pas sensible aux unités influentes ; le second terme représente la somme des impacts individuels de chaque unité sur l'estimateur. Chacun de ces impacts doit donc être borné afin de limiter l'influence de chaque unité et ainsi rendre l'estimateur plus robuste. Nous proposons dans la suite de cette section différentes méthodes permettant de borner cette influence.

3.4.1 Estimation robuste de courbe totale, instant par instant

Démarche de construction d'estimateurs robustes de courbes totales, instant par instant

Pour des variables cibles réelles et non fonctionnelles, Beaumont et al. (2013) proposent une démarche unifiée permettant de construire des estimateurs robustes à l'aide de la notion de biais conditionnels. Leur méthode, que nous allons détailler dans ce paragraphe, consiste à modifier la formule (3.11) de façon à borner les influences conditionnelles de chacune des unités échantillonnées. La manière la plus immédiate de répondre à notre problématique consiste donc à appliquer cette démarche indépendamment sur chacun des instants de discrétisation t_l, $l = 1, \ldots, L$ afin d'estimer la valeur de la courbe totale $t_Y(t_l)$ à cet instant. Les estimations sont alors menées indépendamment sur chaque instant, sans tenir compte du reste de la courbe.

Plus précisément, pour chaque instant $t \in [0, T]$, l'estimateur robuste de la courbe totale est obtenu de l'équation (3.11) en réduisant l'influence de chaque unité de

l'échantillon :

$$\begin{aligned}\hat{\theta}^{\mathrm{R}}(t) &= \left[\hat{\theta}(t) - \sum_{i\in s} \mathrm{B}_{1i}^{\hat{\theta}}(t) - \sum_{i\in \mathrm{U}-s} \mathrm{B}_{0i}^{\hat{\theta}(t)}\right] + \sum_{i\in s} \psi_{c(t)}\left(\mathrm{B}_{1i}^{\hat{\theta}}(t)\right) + \sum_{i\in \mathrm{U}-s} \mathrm{B}_{0i}^{\hat{\theta}}(t) \\ &= \hat{\theta}(t) - \sum_{i\in s} \mathrm{B}_{1i}^{\hat{\theta}}(t) + \sum_{i\in s} \psi_{c(t)}\left(\mathrm{B}_{1i}^{\hat{\theta}}(t)\right) \qquad (3.13)\end{aligned}$$

où $\psi_{c(t)}$ est une fonction de Huber de constante $c(t)$ telle que définie par (3.2).

Les biais conditionnels sont inconnus et doivent être remplacés par leurs estimations comme dans les exemples proposés en (3.7) et (3.9). L'estimateur robuste devient alors

$$\hat{\theta}^{\mathrm{R}}(t) = \hat{\theta}(t) - \sum_{i\in s} \hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t) + \sum_{i\in s} \psi_{c(t)}\left(\hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t)\right), \quad \forall t \in [0, \mathrm{T}]. \qquad (3.14)$$

Le nouvel estimateur robuste $\hat{\theta}^{\mathrm{R}}(t)$ est biaisé mais, potentiellement de variance plus faible que l'estimateur non robuste, c'est pourquoi on espère améliorer la précision globale mesurée par l'erreur quadratique moyenne. Le compromis biais-variance est alors optimisé par le choix de la constante d'ajustement $c(t)$. A mesure que $c(t)$ augmente, l'estimateur se rapproche de l'estimateur non robuste : le biais diminue mais la variance augmente.

Différentes stratégies peuvent être suivies pour choisir la constante optimale $c_{opt}(t)$. Une solution consiste à minimiser un estimateur de l'erreur quadratique moyenne de $\hat{\theta}^{\mathrm{R}}(t)$. Cependant cette approche conduit souvent à des calculs complexes, c'est pourquoi Beaumont et al. (2013) préfèrent utiliser une approche minimax plus facile à implémenter. Cette approche consiste à choisir la constante $c_{opt}(t)$ qui minimise le plus grand biais conditionnel (en valeur absolue) de l'échantillon. On voit donc qu'il s'agit d'une transposition dans le cadre de l'estimation en population finie du principe du minimax présenté plus haut. Plus formellement, la constante optimale $c_{opt}(t)$ est déterminée en cherchant la solution, non nécessairement unique, du problème de minimisation

$$c_{opt}(t) = \arg\min_{c\geq 0} \max_{i\in s} \left|\hat{\mathrm{B}}_{1i}^{\hat{\theta}^{\mathrm{R}}}(t)\right|, \qquad (3.15)$$

où $\hat{\mathrm{B}}_{1i}^{\hat{\theta}^{\mathrm{R}}}(t)$ est l'estimateur du biais conditionnel de l'estimateur robuste. Or l'estimateur robuste s'écrit comme

$$\hat{\theta}^{\mathrm{R}}(t) = \hat{\theta}(t) - \underbrace{\sum_{i\in s} \hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t) + \sum_{i\in s} \psi_{c(t)}\left(\hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t)\right)}_{\Delta(c(t))} \qquad (3.16)$$

et donc, $\mathrm{B}_{1i}^{\hat{\theta}^{\mathrm{R}}}(t) = \mathrm{B}_{1i}^{\hat{\theta}}(t) - \mathrm{E}_p(\Delta(c(t))|\mathrm{I}_i = 1)$. Le biais conditionnel de l'estimateur robuste pour une unité échantillonnée peut donc être estimé par

$$\hat{\mathrm{B}}_{1i}^{\hat{\theta}^{\mathrm{R}}}(t) = \hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t) - \Delta(c(t)). \qquad (3.17)$$

Après quelques calculs, en suivant Beaumont et al. (2013), on obtient

$$\Delta(c_{opt}(t)) = \frac{1}{2}\left(\hat{\mathrm{B}}_{min}^{\hat{\theta}}(t) + \hat{\mathrm{B}}_{max}^{\hat{\theta}}(t)\right), \qquad (3.18)$$

avec $\hat{\mathrm{B}}^{\hat{\theta}}_{min}(t) = \min_{i \in s} \hat{\mathrm{B}}^{\hat{\theta}}_{1i}(t)$ et $\hat{\mathrm{B}}^{\hat{\theta}}_{max}(t) = \max_{i \in s} \hat{\mathrm{B}}^{\hat{\theta}}_{1i}(t)$ respectivement le minimum et le maximum des biais conditionnels estimés $\hat{\mathrm{B}}^{\hat{\theta}}_{1i}(t)$ sur l'échantillon. L'estimateur robuste optimal est alors

$$\hat{\theta}^{\mathrm{R}}_{opt}(t) = \hat{\theta}(t) - \frac{1}{2}\left(\hat{\mathrm{B}}^{\hat{\theta}}_{min}(t) + \hat{\mathrm{B}}^{\hat{\theta}}_{max}(t)\right). \tag{3.19}$$

On remarquera qu'il n'est en fait pas nécessaire d'estimer la constante $c_{opt}(t)$ pour implémenter l'estimateur et qu'il suffit d'estimer les biais conditionnels des unités de l'échantillon puis d'en déterminer le minimum et le maximum.

En particulier, on peut robustifier l'estimateur de Horvitz-Thompson $\hat{t}_{\mathrm{Y}}^{\mathrm{HT,R,1}}$ et par calage $\hat{t}_{\mathrm{Y}}^{cal,\mathrm{R,1}}$, définis respectivement par

$$\hat{t}_{\mathrm{Y}}^{\mathrm{HT,R,1}}(t) = \hat{t}_{\mathrm{Y}}^{\mathrm{HT}}(t) - \frac{1}{2}\left(\hat{\mathrm{B}}^{\hat{t}_{\mathrm{Y}}^{\mathrm{HT}}}_{min}(t) + \hat{\mathrm{B}}^{\hat{t}_{\mathrm{Y}}^{\mathrm{HT}}}_{max}(t)\right) \tag{3.20}$$

et

$$\hat{t}_{\mathrm{Y}}^{cal,\mathrm{R,1}}(t) = \hat{t}_{\mathrm{Y}}^{cal}(t) - \frac{1}{2}\left(\hat{\mathrm{B}}^{\hat{t}_{\mathrm{Y}}^{cal}}_{min}(t) + \hat{\mathrm{B}}^{\hat{t}_{\mathrm{Y}}^{cal}}_{max}(t)\right). \tag{3.21}$$

Illustration sur des données réelles

Afin d'illustrer l'application de cette méthode, voici quelques graphiques réalisés à partir du jeu de données de courbes de charge résidentielles irlandaises que nous présentons au paragraphe 3.6. On observe sur la Figure 3.1 deux exemples de courbes de biais conditionnels fonctionnels estimés, pour un estimateur de Horvitz-Thompson et un échantillon de 100 courbes de charges sélectionné par sondage aléatoire simple : un de ces biais est relativement modéré et le second est plus fort. Ces courbes sont au pas demi-horaire, sur une semaine, et on observe une saisonnalité journalière des biais conditionnels.

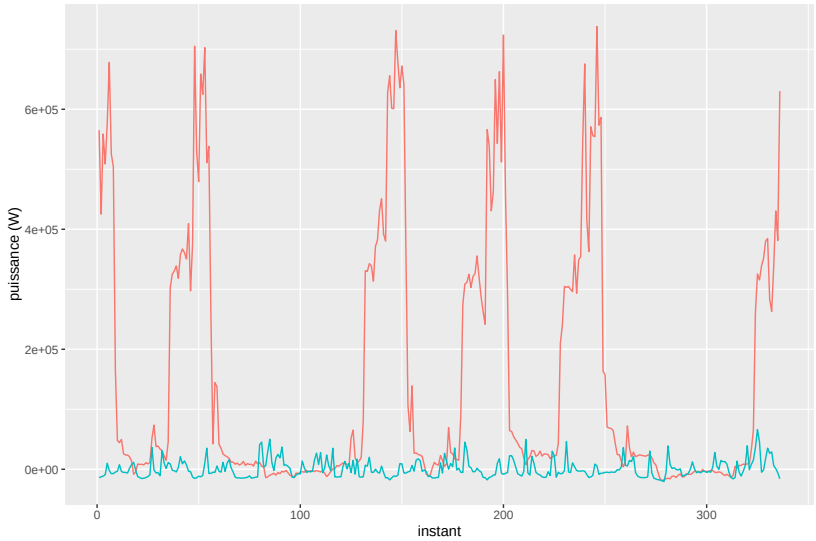

FIGURE 3.1 – Deux biais conditionnels estimés avant troncature, pour un sondage aléatoire simple. Ces biais conditionnels sont des courbes. Ils ont été obtenus par l'équation (3.7) pour H = 1.

Voyons maintenant sur la Figure 3.2, en traits bleus, les limites de troncatures c et $-c$ obtenues en déterminant les constantes $c(t)$ indépendamment sur chaque instant de discrétisation. On représente en traits rouges continus le biais conditionnel tronqué et en traits rouges pointillés le biais conditionnel non tronqué (il s'agit du plus gros des biais conditionnels du graphique précédent) : on constate que les portions de courbes qui dépassent de la zone bleue sont tronquées. On observe donc que les limites de troncature ainsi déterminées sont relativement "plates", même si l'on observe une légère saisonnalité infrajournalière. Elles sont par ailleurs symétriques par construction.

Avantages et inconvénients de la méthode

Cette méthode présente l'avantage indéniable d'être simple à mettre en œuvre. Cependant, on se doute intuitivement qu'elle ne sera pas la plus précise car elle n'exploite pas les corrélations entre les valeurs de la courbe aux différents instants. Or on peut espérer que l'exploitation du caractère fonctionnel de notre problème et des régularités intrinsèques de nos courbes pourrait permettre de gagner en précision. Ceci sera d'ailleurs confirmé empiriquement par nos simulations.

Plus grave encore, le choix des constantes d'ajustement étant réalisé indépendamment sur chaque instant sans tenir compte des autres, cette méthode ne garantit pas la préservation de la structure temporelle de la courbe moyenne ou totale estimée. Cela pourrait alors être extrêmement gênant dans le cadre de certaines applications, par exemple si l'on souhaite ajuster un modèle de prévision sur la courbe moyenne estimée. En effet dans ce cadre les dépendances temporelles de la courbe sont finalement plus importantes que les valeurs de la courbe aux différents instants. De même, il arrive fréquemment qu'on soit plus intéressés par la forme d'une courbe que par son niveau, par exemple pour la construction de profils réglementaires, ce qui implique fi-

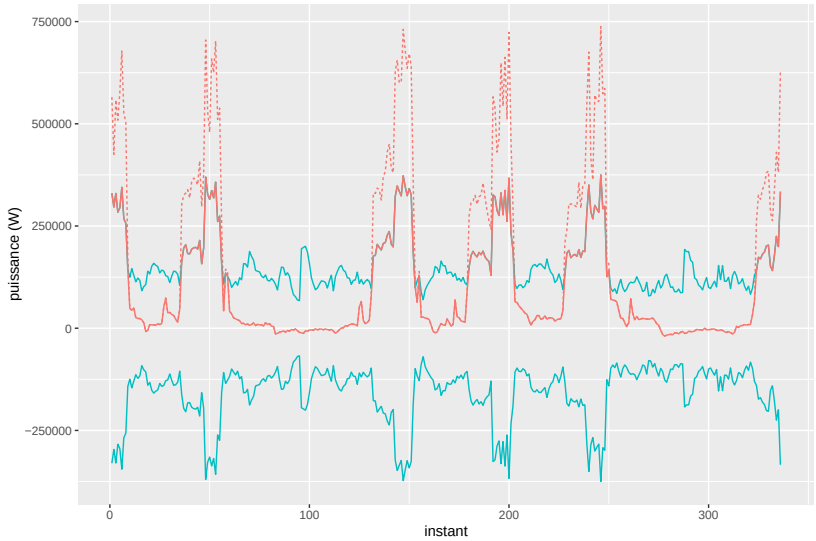

FIGURE 3.2 – Un biais conditionnel estimé (en traits rouge pleins) et la version tronquée (en traits rouge pointillés), pour les limites de troncature (en bleu) définies de façon ponctuelle.

nalement de considérer davantage les relations entre les valeurs aux différents instants que ces valeurs elles-mêmes.

Afin de pallier les défauts que nous venons d'évoquer, nous proposons dans les sous-sections qui suivent deux méthodes conçues avec le souci de préserver la cohérence temporelle de la courbe moyenne estimée. La méthode instant par instant présentée plus haut constitue donc une référence à laquelle on se comparera afin d'évaluer l'apport de la prise en compte de l'aspect fonctionnel de notre problématique. On expose maintenant en détail chacune de ces méthodes.

3.4.2 Estimation robuste sur une base de projection

Dans ce paragraphe, nous proposons de mettre en œuvre une approche en trois temps pour construire nos estimateurs robustes de courbes. La première étape consiste à projeter les courbes dans un espace de dimension finie afin de transformer notre problème fonctionnel en plusieurs sous-problèmes décorrélés d'estimation de totaux ou de moyennes de variables réelles déjà traités dans la littérature. La seconde étape consiste à appliquer les méthodes usuelles univariées - en l'occurrence ici à proposer un estimateur robuste du total des cœfficients en suivant l'approche de Beaumont et al. (2013) présentée plus haut - séparément sur chacun des vecteurs de coordonnées de l'espace de projection. Enfin, dans la troisième étape nous combinons les résultats obtenus précédemment en reconstituant les courbes moyennes ou totales estimées à partir des totaux estimées des cœfficients sur l'espace de projection.

L'utilisation de bases de projection permet donc de préserver la structure de corrélation temporelle de nos données tout en se ramenant à des sous-problèmes d'estimation plus simples que l'on pourra traiter indépendamment.

Comme nous l'avons vu au paragraphe 2.1.2, différentes bases de projection peuvent être choisies, en fonction des caractéristiques du problème traité. Pour des courbes de charge, les bases d'ondelettes sont particulièrement pertinentes. On peut également réaliser une Analyse en Composantes Principales sur les valeurs de la courbe aux différents instants puis travailler dans l'espace des scores de l'ACP. Enfin, on pourrait également réaliser une analyse en Composantes Principales Fonctionnelles en combinant les étapes de projection sur une base fonctionnelle et d'ACP.

Dans la suite de cette sous-section, nous allons considérer le cas de la projection sur l'analyse en Composantes Principales Sphériques, appliquée aux valeurs de la courbe aux différents instants de discrétisation. Comme nous allons le voir, il s'agit d'une version robuste de l'ACP. La démarche serait toutefois similaire sur une autre base de projection.

Nous allons maintenant détailler un peu plus les différentes étapes de notre méthode.

Analyse en Composantes Principales Sphériques

Nos échantillons de courbes de charge contiennent fréquemment des outliers et l'Analyse en Composantes Principales est très sensible aux points atypiques. Par conséquent, on utilisera donc plutôt une version robuste de celle-ci. Différentes propositions d'algorithmes d'ACP robustes existent dans la littérature : nous détaillons ici l'ACP sphérique proposée par Locantore et al. (1999) mais on aurait également pu retenir l'algorithme proposé par Croux and Ruiz-Gazen (2005), qui consiste à déterminer par Projection Pursuit les vecteurs de projection qui préservent un maximum d'information non pas au sens de la variance covariance mais selon un estimateur d'échelle robuste tel que la MAD (Median Absolute Deviation).

Le but de l'ACP sphérique est de construire des estimateurs robustes du centre de l'espace et des vecteurs propres. Cet algorithme a de bonnes propriétés (voir Gervini (2008)) et est aisé à implémenter. Il s'applique à des données multivariées (potentiellement de grande dimension). Comme les points de discrétisation sont équidistants, on peut dans notre cas réaliser une ACP sphérique fonctionnelle en appliquant l'ACP sphérique sur les vecteurs des valeurs des courbes pour les instants de discrétisation.

L'ACP sphérique consiste à considérer les vecteurs propres de la matrice de variance covariance sphéricisée suivante :

$$\Gamma(r,t) = \frac{1}{\mathrm{N}} \sum_{i \in \mathrm{U}} \frac{(\mathrm{Y}_i(r) - m(r))}{||\mathrm{Y}_i - m||} . \frac{(\mathrm{Y}_i(t) - m(t))}{||\mathrm{Y}_i - m||}, \quad r,t \in [0,\mathrm{T}], \tag{3.22}$$

avec m un indicateur robuste de tendance centrale et la norme $||f|| = \left(\int_0^{\mathrm{T}} f^2(t)dt\right)^{1/2}$ usuelle sur $\mathrm{L}^2[0,\mathrm{T}]$.

Intuitivement, en remplaçant la moyenne μ_{Y} par un indicateur robuste de tendance centrale m et les écarts non robustes à la moyenne $\mathrm{Y}_i(t) - \mu_{\mathrm{Y}}(t)$ par les fonctions de norme unitaire $\frac{(\mathrm{Y}_i(t)-m(t))}{||\mathrm{Y}_i - m||}$, on limite l'influence des outliers.

Comme dans Locantore et al. (1999), on choisit comme indicateur de tendance centrale la médiane géométrique (voir Brown (1983)). Dans le contexte de l'estimation en population finie, la courbe médiane de la population composée des éléments

$Y_1, \ldots, Y_N$, dont on suppose qu'elle appartient à $L^2[0,T]$, est définie par

$$m = \arg\min_{y \in L^2[0,T]} \sum_{i \in U} ||Y_i - y||. \tag{3.23}$$

Cette expression constitue une généralisation naturelle de la notion de médiane univariée. Elle est également appelée *médiane spatiale* (Brown (1983)), *médiane* L^1 (Small (1990)) ou encore *médiane géométrique* (Chaudhuri (1996)). Si on suppose que les points $Y_i, \quad i \in U$ ne sont pas concentrés en une ligne, cette médiane existe et est unique (voir Kemperman (1987)). Si $m \neq Y_i$, pour tout $i \in U$, alors elle est l'unique solution de l'équation estimante suivante :

$$\sum_{i \in U} \frac{Y_i - m}{||Y_i - m||} = 0.$$

Elle peut être déterminée par des algorithmes itératifs tels que l'algorithme de Weiszfeld (voir Weiszfeld (1937) et Vardi and Zhang (2000)) pour des données multivariées ou par des algorithmes de gradient (voir Gervini (2008)) pour des données fonctionnelles sparses.

Ensuite, de manière similaire à ce qui est fait pour l'ACP classique, l'Analyse en Composantes Sphériques consiste à déterminer les valeurs propres Λ et vecteurs propres Z correspondants de la matrice de variance-covariance robuste Γ des données projetées, que nous avons définie dans l'équation (3.22). Tout comme pour l'estimateur de tendance centrale, l'influence des outliers sur l'estimation des composantes principales robustes sera alors fortement réduite. En outre, Gervini (2008) montre que si la distribution des Y_i est symétrique, alors la covariance v définie par (2.2) et la covariance sphérique Γ ont les mêmes vecteurs propres orthonormaux.

On peut alors exprimer les courbes Y_i de la population U selon une version robuste de l'expansion de Karhunen-Loève (présentée dans 2.1.3) :

$$Y_i(t) \approx m(t) + \sum_{k=1}^{K} \underbrace{< Y_i - m, Z_k >}_{f_{ik}} Z_k(t), \quad \forall i \in U, \tag{3.24}$$

avec $< f, g >= \int_0^T f(t)g(t)dt$ le produit scalaire usuel sur $L^2[0,T]$. En outre, f_{ik} est la projection de la courbe centrée $Y_i - m$ sur l'espace engendré par la fonction Z_k. L'intérêt de cette décomposition est qu'elle nous permet de décomposer une fonction continue du temps en un nombre fini de problèmes multivariés que nous pourrons résoudre indépendamment les uns des autres à l'aide de techniques de robustification usuelles tout en préservant la structure temporelle de notre problématique.

Comme évoqué dans la remarque 1, nous n'utilisons pas l'ACP dans le but de réduire la dimension de notre problème mais uniquement afin de décomposer le problème en plusieurs sous-problèmes univariés décorrélés ou en tout cas moins corrélés que les problèmes d'origine. On va donc conserver un nombre élevé de composantes K.

Remarque 3. *On remarque que, dans le cas de l'Analyse en Composantes Principales Sphériques, les différentes composantes ne sont pas décorrélées contrairement à l'Analyse en Composantes Principales standard. Néanmoins, on peut s'attendre à ce que les corrélations entre composantes soient plus faibles que les corrélations entre les valeurs de la courbe aux différents instants.*

Estimation de la médiane géométrique et des composantes principales robustes Nous allons maintenant proposer des estimateurs pour la médiane géométrique m ainsi que pour les composantes principales robustes $Z_k, \quad k = 1\ldots, K$. Dans le cadre de l'estimation par sondage en population finie, un estimateur naturel de la médiane géométrique m est la solution $\hat{m}$ de l'équation estimante implicite suivante (voir Chaouch and Goga (2012)),

$$\sum_{i\in s} d_i \frac{Y_i - \hat{m}}{||Y_i - \hat{m}||} = 0. \tag{3.25}$$

On résout cette équation numériquement en quelques itérations d'une version pondérée de l'algorithme de Weiszfeld.

De plus, en suivant Cardot et al. (2010), la matrice de variance-covariance sphéricisée définie par (3.22) peut être estimée par

$$\hat{\Gamma}(r,t) = \frac{1}{N}\sum_{i\in s} d_i \frac{(Y_i(r) - \hat{m}(r))}{||Y_i - \hat{m}||}.\frac{(Y_i(t) - \hat{m}(t))}{||Y_i - \hat{m}||}, \quad r,t \in [0,T], \tag{3.26}$$

avec $\hat{m}$ l'estimateur de la médiane m introduit précédemment.

Finalement, les estimateurs $\hat{Z}_k$ recherchés sont les vecteurs propres de $\hat{\Gamma}$. Dans le paragraphe suivant, nous allons voir comment utiliser l'ACP sphérique pour construire notre estimateur fonctionnel de la courbe moyenne.

Des scores moyens à la courbe moyenne

Un estimateur naturel de la courbe totale de la population t_Y peut être obtenu en sommant les décompositions des courbes selon l'expansion (3.24) et en remplaçant les quantités inconnues par leurs estimateurs :

$$\hat{t}_Y(t) = N\hat{m}(t) + \sum_{k=1}^{K} \hat{F}_k \hat{Z}_k(t), \tag{3.27}$$

avec $\hat{F}_k$ un estimateur de la somme sur la population des scores pour la composante k :

$$F_k = \sum_{i\in U} f_{ik}.$$

On peut utiliser par exemple l'estimateur de Horvitz-Thompson $\hat{F}_k^{HT} = \sum_{i\in s} d_i \hat{f}_{ik}$ ou encore l'estimateur par calage $\hat{F}_k^{cal} = \sum_{i\in s} w_i \hat{f}_{ik}$ avec w_i les poids de calage (voir 2.2.4). Cependant, bien que les estimateurs de médiane et de composantes principales soient robustes, l'estimateur (3.27) ne l'est pas car les estimateurs $\hat{F}_k$ des sommes des coordonnées F_k ne le sont pas.

Nous proposons donc de "robustifier" cet estimateur en rendant robustes les estimateurs $\hat{F}_k, k = 1\ldots K$ selon l'approche de Beaumont et al. (2013) pour des variables réelles présentée dans la sous-section 3.4.1.

Plus précisément, pour chaque composante principale $k = 1,\ldots,K$, la somme sur la population des scores peut être estimée de manière robuste par

$$\hat{F}_k^R = \hat{F}_k - \sum_{i\in s} \hat{B}_{1i}^{\hat{F}_k} + \sum_{i\in s} \Psi_{c_{opt}(k)}(\hat{B}_{1i}^{\hat{F}_k}), \tag{3.28}$$

avec $\hat{B}_{1i}^{\hat{F}_k^{HT}}$ le biais conditionnel de l'unité i pour l'estimateur $\hat{F}_k$. Par exemple, pour l'estimateur de Horvitz-Thompson, on aura

$$\hat{B}_{1i}^{\hat{F}_k} = \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)(\hat{f}_{ik} - \overline{\hat{f}}_{ik,h}), \quad i \in s_h, \tag{3.29}$$

avec $\overline{\hat{f}}_{ik,h}$ la moyenne des scores $\hat{f}_{ik}$ dans la l'échantillon s_h. De plus, $c_{opt}(k)$ est la constante de troncature optimale déterminée par l'approche minimax déjà présentée dans 3.4.1 pour l'estimateur $\hat{F}_k$.

Dans l'équation (3.28), tout comme pour l'estimation instant par instant, il n'est pas forcément nécessaire de déterminer explicitement la constante de troncature optimale $c_{opt}(k)$ mais on peut procéder selon la stratégie définie dans 3.4.1 et construire les estimateurs robustes des moyennes des scores à partir des estimateurs non robustes $\hat{\theta} = \hat{F}_k$ par l'équation (3.19).

Enfin, on en déduit des estimateurs robustes de la courbe totale en remplaçant dans l'équation (3.27) les estimateurs non robustes des scores totaux par leurs versions robustes. Par exemple, pour l'estimateur de Horvitz-Thompson de la courbe totale, on a :

$$\hat{t}_Y^{HT,R,2}(t) = N\hat{m}(t) + \sum_{k=1}^{K} \hat{F}_k^{HT,R}\hat{Z}_k(t). \tag{3.30}$$

De même, pour l'estimateur par calage, on a

$$\hat{t}_Y^{cal,R,2}(t) = N\hat{m}(t) + \sum_{k=1}^{K} \hat{F}_k^{cal,R}\hat{Z}_k(t). \tag{3.31}$$

Remarque 4. *Pour déterminer la constante de troncature sur chaque composante principale, plutôt que d'utiliser le critère minimax, il sera également possible d'employer le nouveau critère que nous proposons dans la sous-section 3.4.4 dans le but d'obtenir des estimateurs d'erreur quadratique moyenne plus précis.*

Remarque 5. *Au lieu d'utiliser une ACP sphérique, on peut projeter les courbes dans une base de fonctions* $\Phi = \phi_1, \ldots, \phi_Q$ *qui engendrent l'espace* $L^2([0,T])$ *des fonctions sur* $[0,T]$ *(par exemple celle des ondelettes). Les courbes* Y *s'écrivent alors*

$$Y_i(t) = \sum_{q=1}^{Q} \alpha_{iq}\phi_q(t) + \epsilon_i(t)$$

où ϵ *est un "résidu d'approximation" de* Y_i *par* $\phi_1 \ldots, \phi_Q$. *En prenant la somme de ces courbes sur la population d'intérêt, on a donc*

$$t_Y(t) = \sum_{q=1}^{Q} \underbrace{\left(\sum_{i \in U} \alpha_{iq}\right)}_{t_{\alpha_q}} \phi_q(t) + \sum_{i \in U} \epsilon_i(t). \tag{3.32}$$

Donc,

$$t_Y(t) \simeq \sum_{q=1}^{Q} t_{\alpha_q}\phi_q(t). \tag{3.33}$$

L'estimateur robuste est alors obtenu en estimant la somme t_{α_q} des cœfficients indépendamment sur chaque vecteur de base par la méthode robuste univariée. Par exemple en utilisant des estimateurs de Horvitz-Thompson, on a alors :

$$\hat{t}_{\mathrm{Y}}^{\mathrm{R,HT,2}}(t) = \sum_{q=1}^{\mathrm{Q}} \hat{t}_{\alpha_q}^{\mathrm{R,HT}} \phi_q(t), \tag{3.34}$$

où

$$\hat{t}_{\alpha_q}^{\mathrm{R,HT}} = \hat{t}_{\alpha_q}^{\mathrm{HT}} - \frac{1}{2}(\hat{\mathrm{B}}_{min}^{\hat{t}_{\alpha_q}^{\mathrm{HT}}} + \hat{\mathrm{B}}_{max}^{\hat{t}_{\alpha_q}^{\mathrm{HT}}}), \tag{3.35}$$

avec $\hat{t}_{\alpha_q}^{\mathrm{HT}} = \sum_{i \in s} d_i \alpha_{iq}$.

Estimation sur la base de l'ACP sphérique : illustration sur des données réelles

Voyons maintenant sur la Figure 3.3 ce que donne l'application de cette méthode sur les mêmes biais conditionnels qu'au paragraphe 3.4.1. Ici, les constantes de troncature ont été déterminées dans l'espace des composnates principales de l'ACP sphérique sur les instants de discrétisation et ce sont les projections des biais conditionnels qui ont été tronquées. Afin de visualiser ce qui se passe, on reconstitue les courbes des biais conditionnels à partir des scores tronqués selon l'équation de Karhunen-Loève (3.24).

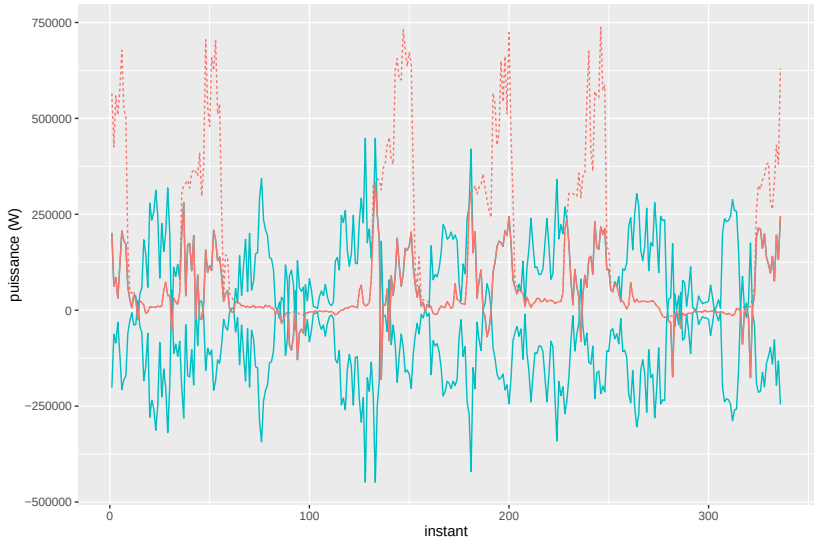

FIGURE 3.3 – Les biais conditionnels tronqués par projection sur la base des composantes principales robustes : on a en bleu les limites déterminées par troncature des biais conditionnels après projection sur les composantes principales, en pointillés rouge un biais conditionnel avant troncature et en traits pleins rouges le biais conditionnel tronqué.

On constate que les limites de troncature ainsi obtenues prennent davantage en compte la saisonnalité infrajournalière des consommations.

3.4.3 Troncature fonctionnelle basée sur la notion de profondeur

Dans ce paragraphe, nous proposons une méthode alternative permettant de réaliser la troncature fonctionnelle des biais conditionnels en tenant compte des corrélations temporelles du problème. Pour cela, nous nous basons sur le concept de profondeur, qui est une mesure d'atypicité pour des données fonctionnelles. Dans le paragraphe suivant, nous définissons plus précisément ce qu'est une profondeur. En particulier, nous présentons les profondeurs que nous avons utilisées dans les tests de la section 3.6. Ensuite, nous expliquons comment se servir de cette notion de profondeur pour tronquer nos biais conditionnels fonctionnels afin de rendre robustes nos estimateurs de courbes totales ou moyennes.

Notion de profondeur pour des données fonctionnelles

Les profondeurs sont des mesures d'atypicité d'un échantillon de données. Elles établissent une relation d'ordre permettant de classer les observations d'un échantillon de l'observation la plus centrale (la moins atypique) à l'observation la plus atypique. Ainsi en dimension 1, cette notion d'observation la plus profonde coïncide avec la médiane. La profondeur s'étend aisément au cadre multivarié : par exemple on peut citer la *half-space depth* de Tukey (1975) qui se base sur le plus petit nombre de points présents dans un demi-espace dont la frontière passe par l'observation considérée.

De nombreuses profondeurs ont été proposées pour des données fonctionnelles dans la littérature. Par exemple, il existe des profondeurs basées sur des distances, telles que la distance de Mahalanobis calculée sur les courbes discrétisées ou encore la profondeur métrique proposée par Gervini (2012) qui considère le rayon de la plus petite boule centrée en l'observation considérée qui contient x% des données. Il existe aussi des profondeurs fonctionnelles construites comme des intégrales de profondeurs instantanées, par exemple celles proposées par Fraiman and Muniz (2001) ou encore Hyndman and Ullah (2007). En outre, Cuesta-Albertos et al. (2006) proposent de projeter les courbes sur des vecteurs choisis aléatoirement puis ensuite de prendre la moyenne des profondeurs obtenues pour chaque projection.

Par la suite, nous testerons deux profondeurs : la Modified Band Depth introduite par López-Pintado and Romo (2009) et la profondeur déduite de la projection des courbes dans l'espace des premières composantes principales sphériques. En effet, ces deux profondeurs ont l'avantage d'être aisées à implémenter et relativement peu gourmandes en temps de calcul. Après avoir présenté ces deux profondeurs, nous expliquerons en détail comment construire des estimateurs robustes en réalisation des troncatures fonctionnelles des biais conditionnels à partir de cette notion de profondeur.

Modified Band Depth

La Modified Band Depth (MBD) est proposée par López-Pintado and Romo (2009). Il s'agit d'une extension de la Band Depth (BD) introduite par les mêmes auteurs. La MBD mesure à quelle fréquence chaque courbe de l'échantillon "est incluse dans l'enveloppe délimitée par un ensemble de ϖ autres courbes" au cours de l'intervalle de temps considéré $[0,T]$. Souvent on considère $\varpi = 2$ courbes. Dans la suite, on conser-

vera cette valeur. Plus la courbe sera fréquemment comprise entre un groupe d'autres courbes, plus sa profondeur sera élevée et plus elle sera considérée comme centrale. Dans notre cas, la profondeur du biais conditionnel d'une unité i appartenant à un échantillon s pour un estimateur $\hat{\theta}$ est définie par

$$\begin{aligned}\mathrm{MBD}_i &= \frac{1}{\binom{2}{n}} \sum_{j,k\in s\, j\neq k} \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \mathbb{1}_{[\min(\hat{\mathrm{B}}^{\hat{\theta}}_{1j}(t_l),\hat{\mathrm{B}}^{\hat{\theta}}_{1k}(t_l))\leq \hat{\mathrm{B}}^{\hat{\theta}}_{1i}(t_l)\leq \max(\hat{\mathrm{B}}^{\hat{\theta}}_{1j}(t_l),\hat{\mathrm{B}}^{\hat{\theta}}_{1k}(t_l))]} \\ &\approx \frac{1}{\binom{2}{n}} \sum_{j,k\in s\, j\neq k} \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \mathbb{1}_{[\min(\hat{\mathrm{B}}^{\hat{\theta}}_{1j}(t),\hat{\mathrm{B}}^{\hat{\theta}}_{1k}(t))\leq \hat{\mathrm{B}}^{\hat{\theta}}_{1i}(t)\leq \max(\hat{\mathrm{B}}^{\hat{\theta}}_{1j}(t),\hat{\mathrm{B}}^{\hat{\theta}}_{1k}(t))]}\, dt.\end{aligned}$$

La Modified Band Depth tient compte de la longueur des intervalles de temps au cours desquels une courbe "sort" de l'enveloppe délimitée par les autres courbes : ainsi une courbe qui sort de l'enveloppe pendant un court intervalle de temps sera considérée comme moins atypique qu'une autre qui en sortirait pendant une plus longue période.

Définition de profondeurs basées sur la projection en dimension finie.

Nous avons également testé un second type de profondeur, mesuré à partir de la projection des courbes dans un espace de dimension finie. On présente le cas de la projection sur la base des composantes principales de l'Analyse en Composantes Principales Sphériques sur les instants de discrétisation mais la généralisation à d'autres espaces de projection est aisée.

Dans ce cadre, on définit la profondeur d'une courbe dans un échantillon comme l'opposé de la distance entre la courbe projetée et le centre de l'espace de projection, c'est-à-dire la somme des cœfficients au carré.

Dans notre cas, on commence par réaliser une ACP sphérique sur l'ensemble des biais conditionnels estimés des unités de l'échantillon. On peut alors exprimer chacun de ces biais conditionnels estimés selon une version robuste de l'équation de Karhunen-Loève, similaire à celle de l'équation (3.24) mais appliquée aux biais conditionnels :

$$\hat{\mathrm{B}}^{\hat{\theta}}_{1i}(t) \approx m_{\hat{\mathrm{B}}}(t) + \sum_{k=1}^{\mathrm{K}_{\hat{\mathrm{B}}}} < \mathrm{Y}_i - m_{\hat{\mathrm{B}}}, \mathrm{Z}_{k,\hat{\mathrm{B}}} > \mathrm{Z}_{k,\hat{\mathrm{B}}}(t), \quad \forall i \in s, \tag{3.36}$$

avec $m_{\hat{\mathrm{B}}}$ la médiane géométrique de l'ensemble des biais conditionnels estimés de l'échantillon et les $\mathrm{Z}_{k,\hat{\mathrm{B}}}$ les composantes principales robustes de l'Analyse en Composantes Principales Sphériques appliqués à ces biais conditionnels estimés. On remplace ces quantités par des estimateurs, déterminés selon l'algorithme décrit dans 3.4.2. La profondeur est ensuite donnée par la somme des carrés des cœfficients de projection :

$$\mathrm{D}_i^{\mathrm{ACP}} = -\sqrt{\sum_{k=1}^{\mathrm{K}_{\hat{\mathrm{B}}}} < \mathrm{Y}_i - \hat{m}_{\mathrm{B}}, \hat{\mathrm{Z}}_{k,\hat{\mathrm{B}}} >^2}, \quad i \in s.$$

Intuitivement, plus cette distance sera petite, plus le biais conditionnel estimé $\hat{\mathrm{B}}^{\hat{\theta}}_{1i}$ sera proche du centre de l'espace de projection et plus il sera considéré comme profond. Ici, contrairement à ce qui est suggéré pour l'estimation par projection sur les

composantes principales, l'ACP est bien utilisée dans une optique de réduction de dimension, et on ne garde qu'un nombre de composantes principales $K_{\hat{B}}$ faible.

Troncature fonctionnelle à partir de la notion de profondeur

Les profondeurs permettent de quantifier le degré d'atypicité d'une courbe au sein d'un jeu de données et d'identifier des outliers fonctionnels. Ce sont ces outliers dont on souhaite borner l'influence lorsque l'on construit des estimateurs robustes. Pour cela, nous allons généraliser l'approche de Beaumont et al. (2013) décrite pour des variables réelles par l'équation (3.14). Plus précisément, nous cherchons à déterminer une fonction de troncature ψ à appliquer aux biais conditionnels estimés afin de les borner. L'estimateur robuste est de la forme (voir formule 3.14) :

$$\hat{\theta}^{R}(t) = \hat{\theta}(t) - \sum_{i \in s} \hat{B}_{1i}^{\hat{\theta}}(t) + \sum_{i \in s} \psi(\hat{B}_{1i}^{\hat{\theta}}(t)), \quad t \in [0, T]. \tag{3.37}$$

La fonction de troncature ψ tient le même rôle que la fonction de Huber ψ_c dans le cas réel mais ici, cette fonction doit s'appliquer à des fonctions. Une idée naturelle serait donc d'utiliser une fonction ψ qui imposerait aux courbes d'être incluses dans une région délimitée par des bornes supérieure et inférieure (fonctionnelles) :

$$\psi_{U,L}(\hat{B}_{1i}^{\hat{\theta}}(t)) = \max\big(\min(\hat{B}_{1i}^{\hat{\theta}}, U(t)), L(t)\big), \forall t \in [0, T].$$

Dans le paragraphe 3.4.1, nous avions déjà employé une fonction de troncature de ce type : $\psi_c = \psi_{U,L}$ avec $U = c$ et $L = -c$. La constante de troncature c était alors déterminée indépendamment pour chaque instant t à l'aide du critère minimax pour des variables réelles. Ici, nous allons chercher à déterminer les limites de troncatures U et L de manière à prendre en compte l'aspect fonctionnel de notre problématique. Pour cela, on déterminera conjointement l'ensemble des valeurs des courbes U et L aux différents instants en appliquant un unique critère fonctionnel. On remarque que, dans le cas général, on n'a pas obligatoirement $L = -U$.

Pour déterminer ces limites de troncature, nous allons procéder de la manière suivante : à partir d'une profondeur judicieusement choisie (par exemple une de celles que nous venons de présenter), on définit une zone "centrale" qui contient intégralement l'ensemble des courbes des biais conditionnels dites "normales" (les "inliers"), définies comme les 50% de courbes les plus profondes, puis nous dilatons cette zone pour obtenir la zone des courbes "non atypiques". Les limites de cette région centrale dilatée sont alors les limites de troncature recherchées U et L : les portions des courbes des biais conditionnels qui ne sont pas contenues dans cette région centrale seront tronquées (c'est-à-dire remplacées par les bornes U et L pour l'intervalle de temps considéré), de la même façon que les valeurs des biais $\hat{B}_{1i}^{\hat{\theta}}$ en dehors de l'intervalle $[-c; c]$ sont remplacées par les bornes c ou $-c$ dans le cas d'une variable réelle.

Cette idée d'utiliser la région centrale qui contient les 50% de courbes les plus profondes et de la dilater d'un facteur fixe a été précédemment utilisée dans le cadre du boxplot fonctionnel proposé par Sun and Genton (2011) et dans le bagplot fonctionnel de Hyndman and Shang (2010) afin de détecter des outliers fonctionnels.

Plus formellement, soit I la région centrale qui contient l'ensemble des 50% des biais conditionnels $\hat{B}_{1i}^{\hat{\theta}}$, $i \in s$ les plus profonds de l'échantillon. La région centrale est la

zone qui contient intégralement ces biais conditionnels. Soit $L_{\hat{B}}$ la borne minimale et $U_{\hat{B}}$ la borne maximale de cette région, pour $t \in [0, T]$:

$$L_{\hat{B}}(t) = \min_{i \in I} \hat{B}^{\hat{\theta}}_{1i}(t), \tag{3.38}$$

$$U_{\hat{B}}(t) = \max_{i \in I} \hat{B}^{\hat{\theta}}_{1i}(t). \tag{3.39}$$

On propose d'utiliser la fonction de troncature suivante :

$$\psi_\alpha(\hat{B}^{\hat{\theta}}_{1i})(t) = \max\Big(\min\Big(\hat{B}^{\hat{\theta}}_{1i}(t), \tilde{m}_{\hat{B}}(t) + \alpha(U_{\hat{B}}(t) - \tilde{m}_{\hat{B}}(t))\Big), \tilde{m}_{\hat{B}}(t) + \alpha(L_{\hat{B}}(t) - \tilde{m}_{\hat{B}}(t))\Big), \tag{3.40}$$

avec $\alpha \geq 0$ un paramètre de dilatation dont la valeur est à déterminer (qui contrôle la taille de la zone des courbes "non atypiques" et donc l'importance de la troncature réalisée) et $\tilde{m}_{\hat{B}}$ la moyenne des biais conditionnels estimés sur l'échantillon (nulle par construction). En pratique, on effectue un lissage sur les limites de troncature $\tilde{m}_{\hat{B}}(t) + \alpha(U_{\hat{B}}(t) - \tilde{m}_{\hat{B}}(t)) = \alpha U_{\hat{B}}(t)$ et $\tilde{m}_{\hat{B}}(t) + \alpha(L_{\hat{B}}(t) - \tilde{m}_{\hat{B}}(t)) = \alpha L_{\hat{B}}(t)$ par une moyenne mobile (d'ordre 5) afin de tirer parti de la structure de corrélations temporelles des données mais aussi d'éviter une troncature trop irrégulière.

Remarque 6. *Pour plus de robustesse, on pourrait également remplacer la courbe des moyennes $\tilde{m}_{\hat{B}}$ par la courbe des médianes instantanées. Cependant des tests préliminaires ont montré que cela détériorait au final la qualité des estimations.*

Le facteur de dilatation $\alpha \geq 0$ joue le même rôle que la constante c dans le cas réel : il assure le compromis entre biais et variance (pour α petit, les courbes sont fortement tronquées, le biais est fort et la variance faible alors que c'est l'inverse pour α grand). Ici, la région centrale dépend donc d'un unique paramètre réel α à la place d'un nombre infini de paramètres réels $c(t), t \in [0, T]$ dans la méthode de troncature instant par instant proposée dans 3.4.1 : la nature fonctionnelle de la problématique est donc prise en compte puisque toutes les limites de troncature sont déterminées simultanément. On peut donc espérer que la troncature préserve davantage la structure temporelle de la courbe moyenne estimée mais aussi que la détermination de la constante de troncature soit plus stable.

En outre, il semble logique de déterminer la valeur optimale α_{opt} du facteur de dilatation par une approche minimax similaire à celle employée dans le cas univarié pour la constante c_{opt} : on choisira donc la valeur α_{opt} qui minimise le maximum de l'intégrale de la valeur absolue des biais conditionnels estimés sur l'échantillon :

$$\alpha_{opt} = \arg\min_{\alpha \geq 0} \left(\max_{i \in s} \sum_{l=1}^{L} |\hat{B}^{\hat{\theta}}_{1i}(t_l) - \Delta_\alpha(t_l)| \right) \tag{3.41}$$

avec

$$\Delta_\alpha(t) = \sum_{j \in s} \Big(\hat{B}^{\hat{\theta}}_{1i}(t) - \psi_\alpha(\hat{B}^{\hat{\theta}}_{1j})(t) \Big). \tag{3.42}$$

Ce nouveau critère peut être vu comme une version fonctionnelle du critère minimax usuel (pour lequel on minimise bien le maximum de la valeur absolue du biais

conditionnel) : ici le biais conditionnel est une courbe, donc on minimise une approximation de sa norme L1.

Cette minimisation peut être réalisée à l'aide d'un algorithme de minimisation numérique [3].

Finalement, l'estimateur robuste de Horvitz-Thompson de la courbe totale avec troncature fonctionnelle est donné par

$$\hat{t}_{\mathrm{Y}}^{\mathrm{HT,R,3}}(t) = \hat{t}_{\mathrm{Y}}^{\mathrm{HT}}(t) - \Delta_{\alpha_{opt}^{\mathrm{HT}}}(t), \quad t \in [0, \mathrm{T}], \tag{3.43}$$

avec

$$\alpha_{opt}^{\mathrm{HT}} = \arg\min_{\alpha \geq 0} \left(\max_{i \in s} \sum_{l=1}^{\mathrm{L}} \left| \hat{\mathrm{B}}_{1i}^{\hat{t}_{\mathrm{Y}}^{\mathrm{HT}}}(t) - \Delta_{\alpha}(t) \right| \right).$$

De même, l'estimateur robuste par calage de la courbe totale avec troncature fonctionnelle est donné par

$$\hat{t}_{\mathrm{Y}}^{cal,\mathrm{R,3}}(t) = \hat{t}_{\mathrm{Y}}^{cal}(t) - \Delta_{\alpha_{opt}^{cal}}(t), \quad t \in [0, \mathrm{T}], \tag{3.44}$$

avec

$$\alpha_{opt}^{cal} = \arg\min_{\alpha \geq 0} \left(\max_{i \in s} \sum_{l=1}^{\mathrm{L}} \left| \hat{\mathrm{B}}_{1i}^{\hat{t}_{\mathrm{Y}}^{cal}}(t_l) - \Delta_{\alpha}(t_l) \right| \right).$$

Illustration sur des données réelles

Voyons sur la Figure 3.4 ce que donne l'application de cette méthode sur les biais conditionnels déjà présentés dans la Figure 3.1. Ici on applique la troncature fonctionnelle basée sur la Modified Band Depth. Les limites U et L de la région centrale sont en vert, la moyenne m est nulle par construction. Comme précédemment, les traits bleus représentent les limites de la région de troncature obtenues en dilatant la région centrale d'un facteur α. Les lignes pointillées rouges représentent un biais conditionnel avant troncature et les lignes continues rouges le biais conditionnel tronqué.

On remarque ici que les limites de la région de troncature reflètent beaucoup plus naturellement les saisonnalités des courbes et donc des biais conditionnels. Par ailleurs, la région de troncature est maintenant non symétrique, et beaucoup plus large.

Pour visualiser de façon plus nette la différence entre les bandes de troncatures délimitées par chacune des méthodes, on regroupe finalement les trois régions centrales sur la même Figure 3.5.

3. en R, on utilise la fonction stats : :optimize, qui réalise une combinaison de *golden section search* et d'interpolations paraboliques successives, voir Brent (2013)

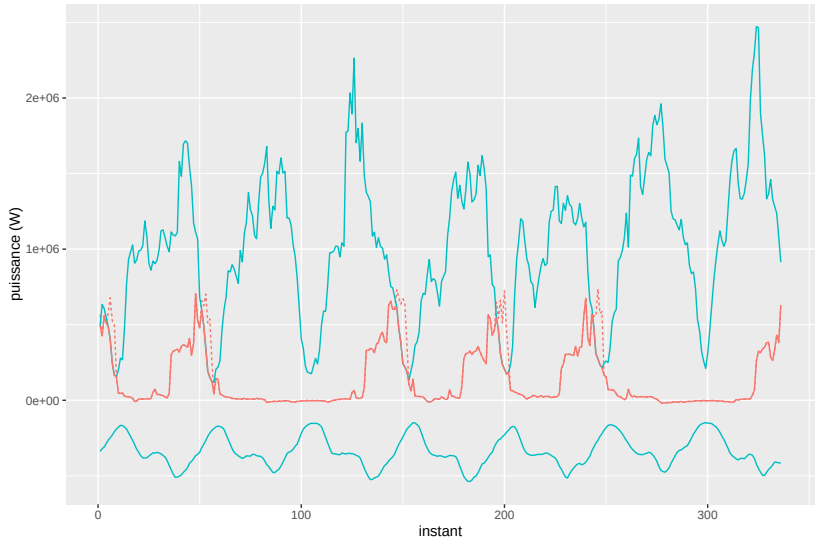

FIGURE 3.4 – Biais conditionnels tronqués par la méthode fonctionnelle basée sur la Modified Band Depth. Comme précédemment, ce graphique montre les limites de la zone de troncature fonctionnelle en bleu, un exemple de biais conditionnel en pointillés rouges et ce biais conditionnel après troncature en traits pleins rouges.

FIGURE 3.5 – Comparaison des bornes de troncature obtenues par les différentes méthodes.

3.4.4 Proposition de nouveau critère pour le choix des constantes d'ajustement

Pour les trois approches proposées dans la sous-section précédente, la sélection de la constante de troncature optimale est effectuée selon des critères de type minimax. L'estimateur robuste repose donc sur des fonctions non lisses de la mesure (minimum

et maximum). Par conséquent il est délicat d'estimer par bootstrap l'erreur quadratique moyenne de l'estimateur robuste ainsi construit. Dans cette sous-section, on propose donc, pour chacune des approches présentées ci-dessus, un critère alternatif basé sur une fonction plus lisse des indicatrices d'appartenance des unités à l'échantillon dont on espère qu'il permettra des estimations d'erreur quadratique moyenne plus réalistes. Ce critère peut se substituer au critère minimax pour le choix de la constante de troncature optimale c_{opt} ou α_{opt}. Le reste de la méthode de construction des estimateurs robustes reste alors inchangé.

Nouveau critère pour l'estimation robuste instant par instant

On se place dans le cadre de l'estimation robuste de courbe totale instant par instant de la sous-section 3.4.1. L'expression de l'estimateur robuste $\hat{\theta}^{R}$ reste celle définie par (3.14) et, pour chaque instant $t \in [0, T]$, on cherche à déterminer la constante $c(t)$ optimale, que nous nommerons $c_{opt}^{alt}(t)$, selon un critère alternatif.

Plus précisément, on choisira la constante de troncature $c_{opt}^{alt}(t)$ telle que

$$c_{opt}^{alt}(t) = \arg\min_{c \geq 0} \sum_{i \in s} \left| \hat{B}_{1i}^{\hat{\theta}}(t) - \Delta(c(t)) \right|^{q}, \quad t \in [0, T], \tag{3.45}$$

où $\Delta(c(t)) = \sum_{i \in s} \hat{B}_{1i}^{\hat{\theta}}(t) - \sum_{i \in s} \psi_{c(t)} \left(\hat{B}_{1i}^{\hat{\theta}}(t) \right)$ et q un entier à choisir. Quand q augmente, l'estimateur robuste basé sur ce critère se rapproche de l'estimateur minimax défini dans 3.43.

Cette minimisation, comme toutes celles de cette sous-section sur le nouveau critère, pourra être réalisée numériquement par un algorithme standard, par exemple celui de Newton-Raphson.

L'idée sous-jacente de ce critère est comme pour le minimax de pénaliser les biais conditionnels trop élevés. Cependant ce nouveau critère ne prend plus en compte uniquement le "pire des biais conditionnels" comme dans l'approche minimax mais la somme sur l'échantillon des biais conditionnels élevés à la puissance q. Ainsi, le fait qu'une unité appartienne ou non à l'échantillon n'engendre alors plus un changement brusque dans la valeur de la constante comme cela pouvait être le cas pour l'approche minimax. En effet, pour celle-ci, la présence ou l'absence dans l'échantillon des individus dont le biais conditionnel est minimal ou maximal engendre des discontinuités dans la valeur de la constante de troncature et donc dans la valeur de l'estimateur. On peut alors espérer que, puisque l'estimateur robuste est une fonction "plus lisse" de la mesure, sa variance puisse être approximée plus précisément par bootstrap.

L'entier q, qui est la puissance à laquelle on élève les biais, permettra de choisir à quel point on privilégie le fait de réduire plutôt les "pires" biais conditionnels individuels plutôt que de borner globalement l'ensemble des biais conditionnels : plus la puissance q sera grande, plus on donnera d'importance dans le critère aux valeurs extrêmes du biais conditionnel et donc plus on tendra vers le critère minimax. En employant une puissance q élevée, on obtient donc un estimateur proche de l'estimateur minimax, mais possédant de meilleures propriétés de régularité dont on espère qu'elles permettront d'estimer de manière plus précise l'erreur quadratique moyenne par bootstrap.

Nouveau critère pour l'estimation robuste sur une base de projection

On se place maintenant dans le cadre de l'estimation robuste de courbe totale par projection sur la base des composantes principales sphériques (ou, plus généralement, par projection dans une base de fonctions). Les expressions des estimateurs robustes de Horvitz-Thompson et par calage restent celles définies respectivement par (3.30) et (3.31), et l'expression des estimateurs des totaux des scores sur chaque composante principale robuste reste celle définie par (3.28). On cherche cette fois à déterminer la constante optimale pour chaque composante principale par le critère alternatif, que nous nommerons ici $c_{opt}^{alt}(k), \quad k = 1, \ldots, \mathrm{K}$. Le principe de la méthode est le même que pour la troncature instant par instant, et la constante optimale est définie comme :

$$c_{opt}^{alt}(k) = \arg\min_{c \geq 0} \sum_{i \in s} \left| \hat{\mathrm{B}}_{1i}^{\hat{\mathrm{F}}_k} - \Delta(c(k)) \right|^q, \quad t \in [0, \mathrm{T}], \tag{3.46}$$

où $\Delta(c(k)) = \sum_{i \in s} \hat{\mathrm{B}}_{1i}^{\hat{\mathrm{F}}_k} - \sum_{i \in s} \psi_{c(k)}(\hat{\mathrm{B}}_{1i}^{\hat{\mathrm{F}}_k})$ et q est un entier à choisir.

Nouveau critère pour l'estimation robuste par troncature fonctionnelle

On se place maintenant dans le cadre de l'estimation robuste de courbe totale ou moyenne par troncature fonctionnelle des biais conditionnels, et l'expression de l'estimateur robuste reste celle définie par l'équation (3.37). On cherche à déterminer par le critère alternatif la constante optimale, que nous nommerons ici α_{opt}^{alt}.

De manière très similaire à ce qui a été fait précédemment pour des variables réelles, le problème de minimisation, défini pour le critère minimax par (3.41) devient pour ce nouveau critère :

$$\alpha_{opt}^{alt} = \arg\min_{\alpha \geq 0} \left(\sum_{i \in s} \sum_{l=1}^{\mathrm{L}} \left| \hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t_l) - \Delta_\alpha(t_l) \right|^q \right), \tag{3.47}$$

où

$$\Delta_\alpha(t) = \sum_{j \in s} \left(\hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t) - \psi_\alpha(\hat{\mathrm{B}}_{1j}^{\hat{\theta}})(t) \right) \tag{3.48}$$

et q un entier à choisir.

3.5 Estimation d'erreur quadratique moyenne (EQM) pour les estimateurs robustes de courbes

Dans cette section, on définit des estimateurs ponctuels de l'Erreur Quadratique Moyenne de nos estimateurs robustes. L'erreur quadratique moyenne est préférée à la variance car elle permet de prendre en compte le biais des estimateurs robustes. Son expression est définie par l'équation (2.7). On s'intéresse ici au cas particulier d'un estimateur de Horvitz-Thompson de total $\hat{\theta} = \hat{t}_Y^{HT}$ mais la démarche est similaire notamment pour un estimateur de moyenne et/ou un estimateur par calage.

Afin d'estimer l'erreur quadratique moyenne de leur estimateur $\hat{\theta}^R$, Beaumont et al. (2013) proposent d'utiliser un estimateur de la forme proposée par Gwet and Rivest (1992) (exprimée ici pour un instant t donné de $[0, T]$) :

$$\hat{\text{EQM}}(\hat{\theta}^R(t)) = v_p\left(\hat{\theta}^R(t)\right) + \max\left(0, \left(\hat{\theta}^R(t) - \hat{\theta}(t)\right)^2 - v_p\left(\hat{\theta}^R(t) - \hat{\theta}(t)\right)\right), \tag{3.49}$$

où $v_p\left(\hat{\theta}^R(t)\right)$ et $v_p\left(\hat{\theta}^R(t) - \hat{\theta}(t)\right)$ sont des estimateurs convergents de la variance sous le plan respectivement de $\mathbb{V}\left(\hat{\theta}^R(t)\right)$ et $\mathbb{V}\left(\hat{\theta}^R(t) - \hat{\theta}(t)\right)$.

Dans la suite de cette section, nous montrons comment décliner cette approche dans le contexte des données fonctionnelles, pour les trois approches proposées (instant par instant, projection sur une base de dimension finie et troncature fonctionnelle). Pour l'ensemble de ces approches, trois types de méthodes d'estimation d'erreur quadratique moyenne sont proposés : les estimateurs explicites, pour lesquels on suppose comme dans Gwet and Rivest (1992) que la ou les constantes de troncature sont fixées, les bootstraps populationnels et les bootstraps généralisés.

Dans les deux derniers cas, on propose de ne pas considérer la constante comme fixe mais de la réestimer dans chaque rééchantillon selon la stratégie décrite dans 3.5.2 et 3.5.3. En effet, le fait de considérer comme déterministe une constante qui dépend en réalité de l'échantillon revient à omettre une source de variabilité et donc à sous-estimer l'erreur commise, comme le montreront les tests. Les estimateurs robustes obtenus en choisissant les constantes de troncature optimales par notre nouveau critère proposé au paragraphe 3.4.4 sont "bootstrappables". En revanche, ceux dont la constante est déterminée par le critère minimax ne le sont pas car ils font intervenir des minima et des maxima sur l'échantillon. On propose donc pour approximer les erreurs quadratiques moyennes de ceux-ci par bootstrap d'utiliser un critère approximé pour déterminer la constante optimale dans les réplications.

3.5.1 Estimateurs explicites d'erreur quadratique moyenne

Dans ce paragraphe, on suppose qu'on dispose d'un estimateur explicite convergent de variance sous le plan v_p . Par exemple, dans notre cas où l'estimateur considéré $\hat{\theta}$ est l'estimateur de Horvitz-Thompson du total, on pourra se rapporter à l'équation (2.10). Comme suggéré dans Beaumont et al. (2013), les constantes de troncature sont considérées comme fixées. On détaille la démarche pour chacune de nos trois approches de construction d'estimateurs robustes.

Estimateurs explicites d'EQM pour l'estimation robuste instant par instant

Ici, on suppose qu'on a construit nos estimateurs robustes en appliquant la méthode d'estimation robuste pour des totaux réels sur les instants de discrétisation comme proposé dans 3.4.1. Soit $c_{opt}(t)$ la constante optimale pour chaque instant $t \in [0,\mathrm{T}]$, solution du problème de minimisation (3.15) (ou du problème (3.45) si on choisit d'utiliser notre nouveau critère plutôt que le critère minimax). On va considérer cette constante comme fixée. Posons

$$z_i^{opt}(t) = \pi_i \left(\psi_{c_{opt}(t)}\left(\hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t)\right) - \hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t)\right), \quad \forall i \in s, \quad \forall t \in [0,\mathrm{T}]$$

et

$$\mathrm{Y}_i^*(t) = \mathrm{Y}_i(t) + z_i^{opt}(t), \quad \forall i \in s, \quad \forall t \in [0,\mathrm{T}].$$

Cette dernière quantité peut être vue comme la version "tronquée" de Y_i. Ici, comme on a considéré $c_{opt}(t)$ comme fixée, $\mathrm{Y}_i^*(t)$ est également fixée.

On remarque alors que $\hat{\theta}^{\mathrm{R}}(t)$ s'exprime comme l'estimateur de Horvitz-Thompson du total sur la population de $\mathrm{Y}_i^*(t)$:

$$\hat{\theta}^{\mathrm{R}}(t) = \sum_{i \in s} \frac{1}{\pi_i} \mathrm{Y}_i^*(t). \tag{3.50}$$

De même, $\hat{\theta}^{\mathrm{R}}(t) - \hat{\theta}(t)$ s'exprime comme l'estimateur de Horvitz-Thompson du total de $z_i^{opt}(t)$ sur la population. On peut donc proposer des estimateurs des variances de ces quantités :

$$v_p\left(\hat{\theta}^{\mathrm{R}}(t)\right) = v_p\left(\sum_{i \in s} d_i \mathrm{Y}_i^*(t)\right) = \sum_{i \in s}\sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\mathrm{Y}_i^*}{\pi_i} \frac{\mathrm{Y}_j^*}{\pi_j}. \tag{3.51}$$

$$v_p\left(\hat{\theta}^{\mathrm{R}}(t) - \hat{\theta}(t)\right) = v_p\left(\sum_{i \in s} d_i z_i^{opt}\right) = \sum_{i \in s}\sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{z_i^{opt}}{\pi_i} \frac{z_j^{opt}}{\pi_j}. \tag{3.52}$$

Il suffit ensuite de réinjecter ces expressions dans l'équation (3.49) pour obtenir une approximation de l'erreur quadratique moyenne.

Remarque 7. *En pratique, pour le critère minimax nous n'avons pas eu besoin de calculer les constantes $c_{opt}(t), \quad t \in [0,\mathrm{T}]$ pour construire l'estimateur robuste puisque nous pouvions utiliser directement l'expression (3.19). Cependant, pour approximer l'erreur quadratique moyenne il est nécessaire de les estimer explicitement.*

Estimateurs explicites d'EQM pour la troncature dans l'espace des composantes principales

Dans ce paragraphe, on procède de manière similaire à ce qui était fait précédemment. Soient $c_{opt}(k), \quad k \in 1,\dots,\mathrm{K}$, les constantes optimales de troncature pour chaque composante principale. On va considérer ces constantes comme fixées. On définit ensuite

$$z_{ik}^{opt} = \pi_i \left(\psi_{c_{opt}(k)}(\hat{\mathrm{B}}_{1i}^{\hat{\mathrm{F}}_k^{\mathrm{HT}}}) - \hat{\mathrm{B}}_{1i}^{\hat{\mathrm{F}}_k^{\mathrm{HT}}}\right), \quad \forall i \in s, \quad k \in 1,\dots,\mathrm{K}$$

avec $\hat{F}_k^{HT}$ défini dans 3.4.2, et

$$f_{ik}^* = f_{ik} + z_{ik}^{opt}, \quad \forall i \in s.$$

Cette dernière expression peut être vue comme la version tronquée du score de l'unité i pour la composante k. On déduit de l'expansion de Karhunen-Loève (3.30) l'expression suivante de l'estimateur robuste :

$$\hat{\theta}^R(t) = N\hat{m}(t) + \sum_{k=1}^{K} \sum_{i \in s} \frac{f_{ik}^*}{\pi_i} \hat{Z}_k(t), \quad \forall t \in [0, T]. \tag{3.53}$$

Notre objectif est de proposer des approximations de $\mathbb{V}\left(\hat{\theta}^R(t)\right)$ et $\mathbb{V}\left(\hat{\theta}^R(t) - \hat{\theta}(t)\right)$. Pour cela, nous allons utiliser la technique de la linéarisation présentée dans 2.2.5.

Tout d'abord, en suivant Chaouch and Goga (2012), on obtient l'expression de la linéarisée de $\hat{m}$ l'estimateur de la médiane fonctionnelle :

$$\mathbb{V}_p(\hat{m}(t)) \approx \mathbb{V}_p\left(\sum_{i \in s} \frac{u_i(t)}{\pi_i}\right), \quad t \in [0, T], \tag{3.54}$$

avec $u_i = \Gamma^{-1}(\frac{Y_i - m}{||Y_i - m||})$, où Γ est l'opérateur de covariance pour les données "sphéricisées" défini par (3.22). Cette linéarisée dépend de quantités inconnues m et Γ . On peut cependant l'estimer en remplaçant celles-ci par leurs estimateurs :

$$\hat{u}_i = N\hat{\Gamma}^{-1}\left(\frac{Y_i - \hat{m}}{||Y_i - \hat{m}||}\right),$$

avec $\hat{\Gamma}$ défini dans (3.26).

En outre, on a

$$\sum_{k=1}^{K} \left(\hat{F}_k^R - F_k\right) Z_k(t) \approx \sum_{k=1}^{K} \left(\sum_{i \in s} d_i f_{ik}^*\right) Z_k(t) = \sum_{i \in s} \frac{1}{\pi_i} \left(\sum_{k=1}^{K} f_{ik}^* Z_k(t)\right), \quad t \in [0, T].$$

donc

$$\hat{\theta}^R(t) - \theta(t) \approx N\hat{m}(t) - Nm(t) + \sum_{i \in s} \frac{1}{\pi_i} \left(u_i(t) + \sum_{k=1}^{K} f_{ik}^* Z_k(t)\right), \quad t \in [0, T].$$

La variance de cette quantité peut donc être approximée, pour tout $t \in [0, T]$ par

$$\mathbb{V}_p(\hat{\theta}^R(t) - \theta(t)) \approx \mathbb{V}\left(\sum_{i \in s} \frac{1}{\pi_i} \left(u_i(t) + \sum_{k=1}^{K} f_{ik}^* Z_k(t)\right)\right) \tag{3.55}$$

$$\approx \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{u_i(t) + \sum_{k=1}^{K} f_{ik}^* Z_k(t)}{\pi_i} \frac{u_j(t) + \sum_{k=1}^{K} f_{jk}^* Z_k(t)}{\pi_j}. \tag{3.56}$$

En remplaçant les différentes quantités inconnues par leurs estimateurs, on obtient finalement l'estimateur souhaité, pour tout $t \in [0, T]$:

$$v_p\left(\hat{\theta}^R(t) - \theta(t)\right) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\hat{u}_i(t) + \sum_{k=1}^{K} f_{ik}^* \hat{Z}_k(t)}{\pi_i} \frac{\hat{u}_j(t) + \sum_{k=1}^{K} f_{jk}^* \hat{Z}_k(t)}{\pi_j}. \tag{3.57}$$

Par ailleurs, on a

$$\hat{\theta}^{\mathrm{R}}(t) - \hat{\theta}(t) = \sum_{k=1}^{\mathrm{K}} \left(\sum_{i \in s} d_i z_{ik}^{opt} \right) \hat{\mathrm{Z}}_k(t), \quad t \in [0, \mathrm{T}]. \tag{3.58}$$

La variance de cette quantité peut donc être approximée par

$$\mathbb{V}_p\left(\hat{\theta}^{\mathrm{R}}(t) - \theta(t)\right) \approx \mathbb{V}_p\left(\sum_{i \in s} \frac{1}{\pi_i} \left(\sum_{k=1}^{\mathrm{K}} z_{ik}^{opt} \hat{\mathrm{Z}}_k(t) \right) \right), \quad t \in [0, \mathrm{T}]. \tag{3.59}$$

Elle peut être estimée par :

$$v_p\left(\hat{\theta}^{\mathrm{R}}(t) - \theta(t)\right) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\sum_{k=1}^{\mathrm{K}} z_{ik}^{opt} \hat{\mathrm{Z}}_k(t)}{\pi_i} \frac{\sum_{k=1}^{\mathrm{K}} z_{jk}^{opt} \hat{\mathrm{Z}}_k(t)}{\pi_j}, \quad t \in [0, \mathrm{T}]. \tag{3.60}$$

En insérant les expressions (3.57) et (3.60) dans l'équation (3.49), on obtient finalement l'estimateur d'erreur quadratique moyenne attendu.

Estimateurs explicites d'EQM pour la troncature fonctionnelle

Pour la méthode de construction d'estimateurs robustes par troncature fonctionnelle présentée au paragraphe 3.4.3, on procède de la même manière que dans les paragraphes précédents. Plus précisément, soit α_{opt} la constante de troncature optimale, solution du problème de minimisation (3.41) (ou du problème (3.47) si on choisit d'utiliser notre nouveau critère plutôt que le critère minimax). On va considérer cette constante comme fixée. On pose

$$z_i^{opt}(t) = \pi_i \left(\psi_{\alpha_{opt}}(\hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t)) - \hat{\mathrm{B}}_{1i}^{\hat{\theta}}(t) \right), \quad \forall i \in s, \quad \forall t \in [0, \mathrm{T}],$$

avec ψ_α définie par (3.40). Enfin, on pose

$$\mathrm{Y}_i^*(t) = \mathrm{Y}_i(t) + z_i^{opt}(t), \quad \forall i \in s, \quad \forall t \in [0, \mathrm{T}].$$

Cette dernière quantité peut être vue comme la version "tronquée" de Y_i.

Comme dans 3.5.1, on remarque que $\hat{\theta}^{\mathrm{R}}(t)$ s'exprime comme l'estimateur de Horvitz-Thompson du total sur la population de $\mathrm{Y}_i^*(t)$. De même, $\hat{\theta}^{\mathrm{R}}(t) - \hat{\theta}(t)$ s'exprime comme l'estimateur de Horvitz-Thompson du total sur la population de $z_i^{opt}(t)$. Des estimations des variances de ces estimateurs peuvent donc être fournies par l'équation (3.51) et peuvent ensuite être réinjectées dans (3.49) pour obtenir une approximation de l'erreur quadratique moyenne.

3.5.2 Estimateurs d'EQM par bootstrap populationnel

Dans la sous-section précédente, pour les trois approches proposées, nous avons considéré les constantes de troncature comme fixes. Toutefois en réalité, celles-ci dépendent de l'échantillon et on omet donc une source de variabilité qui pourrait potentiellement engendrer une sous-estimation de l'incertitude. Pour pallier ce défaut, on

propose donc dans cette sous-section et la suivante de recourir à des techniques de bootstrap, en considérant des constantes de troncature choisies à partir des données pour chaque rééchantillon.

Ces méthodes par réplication sont toutefois sensiblement plus gourmandes en temps de calcul que l'utilisation d'estimateurs explicites d'erreurs quadratiques moyennes, elles ne sont donc à privilégier que si on suppose que la sous-estimation engendrée par le fait de fixer la constante est importante ou encore pour les estimateurs complexes pour lesquels on ne dispose pas d'estimateurs de variance explicite.

Dans ce paragraphe, nous mettons en œuvre un bootstrap populationnel pour construire les estimateurs de variance $v_p\left(\hat{\theta}^{\mathrm{R}}(t)-\theta(t)\right)$ et $v_p\left(\hat{\theta}^{\mathrm{R}}(t)\right), \quad t \in [0,\mathrm{T}]$. Les estimateurs de variance ainsi obtenus sont ensuite réinjectés dans l'expression de l'erreur quadratique moyenne (3.49). La démarche est identique pour les trois approches de construction d'estimateur robuste de courbe moyenne ou totale.

Pour les trois approches de construction d'estimateurs robustes proposées, les estimateurs dont la constante de troncature est déterminée suivant les nouveaux critères proposés dans la section 3.4.4 sont par construction des fonctions lisses de la mesure de sondage et on peut donc s'attendre à ce que le bootstrap donne des résultats satisfaisants. En revanche, les estimateurs robustes dont la constante est sélectionnée par le critère minimax, qui se basent sur des fonctions non lisses de la mesure de sondage (minimum et maximum sur l'échantillon) ne sont pas bootstrappables (voir les conditions de régularité énoncées dans Bertail and Combris (1997)). On suggère donc dans ce cas de remplacer le minimum et le maximum respectivement par les quantiles à 1% et à 99 % sur l'échantillon.

On définit les formules d'approximation de variance par bootstrap, pour un grand nombre B de réplications :

$$v_p\left(\hat{\theta}^{\mathrm{R}}(t)\right)=\frac{1}{\mathrm{B}-1}\sum_{b=1}^{\mathrm{B}}\left[\hat{\theta}^{*\mathrm{R}(b)}(t)-\overline{\hat{\theta}^{*\mathrm{R}}}(t)\right]^2 \tag{3.61}$$

et

$$v\left(\hat{\theta}^{\mathrm{R}}(t)-\hat{\theta}(t)\right)=\frac{1}{\mathrm{B}-1}\sum_{b=1}^{\mathrm{B}}\left[\left(\hat{\theta}^{*\mathrm{R}(b)}(t)-\hat{\theta}^{*(b)}(t)\right)-\left(\overline{(\hat{\theta}^{*\mathrm{R}}}(t)-\overline{(\hat{\theta}^{*}}(t)\right)\right]^2, \tag{3.62}$$

avec

$$\overline{\hat{\theta}^{*\mathrm{R}}}(t)=\frac{1}{\mathrm{B}}\sum_{b=1}^{\mathrm{B}}\hat{\theta}^{*\mathrm{R}(b)}(t) \tag{3.63}$$

$$\overline{\hat{\theta}^{*}}(t)=\frac{1}{\mathrm{B}}\sum_{b=1}^{\mathrm{B}}\hat{\theta}^{*(b)}(t). \tag{3.64}$$

Le bootstrap populationnel a été proposé par Booth et al. (1994) et nous l'avons présenté dans le paragraphe 2.2.5. Dans notre contexte d'estimation robuste fonctionnelle, il est implémenté selon l'algorithme 1.

3.5.3 Estimateurs d'EQM par bootstrap généralisé

Dans ce dernier paragraphe, nous proposons une méthode d'estimation d'erreur quadratique moyenne basée sur un autre type de bootstrap : le bootstrap généralisé,

Soit B le nombre de réplications, (B grand).

pour $b \leftarrow 1$ ***à*** B **faire**

Générer une superpopulation U^* en répliquant chaque unité i $w_i = [\frac{1}{\pi_i}]$ fois, où $[x]$ la partie entière de x puis en tirant un complément de taille $n' = N - \sum_{i \in s}[\frac{1}{\pi_i}]$ par sondage aléatoire simple sans remise dans U de façon à ce que la taille de U^* soit égale à la taille de U.

Tirer un rééchantillon s^* parmi U^* selon le plan p qui a conduit à la sélection de s.

Calculer l'estimateur non robuste $\hat{\theta}^{*(b)}$ et l'estimateur robuste $\hat{\theta}^{*R(b)}$ à partir de l'échantillon s^* . En particulier, les constantes de troncature devront être estimées à nouveau sur l'échantillon s^* à partir des données de ce rééchantillon.

fin

A partir des estimateurs $\theta^{*R(b)}$ et $\theta^{*(b)}$ obtenus à chaque itération $b = 1, \ldots B$, calculer les expressions des estimateurs de variance $v_p\left(\theta^R(t)\right)$ et $v_p\left(\hat{\theta}^R(t) - \hat{\theta}(t)\right)$ par les équations (3.61) et (3.62).

retourner $v_p\left(\hat{\theta}^R(t)\right)$ *et* $v_p\left(\hat{\theta}^R(t) - \hat{\theta}(t)\right)$

Algorithme 1 : Calcul des variances de $v_p\left(\hat{\theta}^R(t)\right)$ et $v_p\left(\hat{\theta}^R(t) - \hat{\theta}(t)\right)$ par bootstrap populationnel.

proposé par Bertail and Combris (1997). La démarche d'estimation d'erreur quadratique moyenne sera identique à celle basée sur le bootstrap populationnel, et seule la méthode de construction des estimateurs répliqués $\hat{\theta}^{*(b)}$ et $\hat{\theta}^{*R(b)}$ diffère. En effet, dans le bootstrap généralisé, l'échantillon employé pour l'estimation reste le même à chaque réplication, mais c'est le poids attribué à chacune des unités de l'échantillon qui change. Ces poids seront notés $W_i^{*(b)}, \quad i = 1, \ldots, n, \quad b = 1, \ldots, B$. Dans notre contexte d'estimation robuste de courbe totale ou moyenne, le bootstrap populationnel est implémenté suivant l'algorithme 2. Nous détaillons par la suite les expressions des estimateurs non robustes et robustes dans les réplications pour les différentes approches dans le cas de l'estimateur de Horvitz-Thompson du total.

Soit B le nombre de réplications, (B grand).
pour $b \leftarrow 1$ ***à*** B **faire**
Simuler les poids de réplication $(\mathrm{W}_i^{*(b)})_{i=1}^n$ selon une loi normale dont les moments sont ceux définis dans la sous-section 2.2.5.
En déduire l'estimateur non robuste $\hat{\theta}^{*(b)}$ ainsi que les biais conditionnels $\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}$ et enfin les estimateurs robustes $\hat{\theta}^{*\mathrm{R}(b)}$ en appliquant la démarche d'estimation robuste sur le rééchantillon.
fin
A partir des estimateurs $\hat{\theta}^{*\mathrm{R}(b)}$ et $\hat{\theta}^{*(b)}$ obtenus à chaque itération $b = 1, \ldots \mathrm{B}$, calculer les expressions des estimateurs de variance $v_p\left(\hat{\theta}^{\mathrm{R}}(t)\right)$ et $v_p\left(\hat{\theta}^{\mathrm{R}}(t) - \hat{\theta}(t)\right)$ par les équations (3.61) et (3.62).
retourner $v_p\left(\hat{\theta}^{\mathrm{R}}(t)\right)$ *et* $v_p\left(\hat{\theta}^{\mathrm{R}}(t) - \hat{\theta}(t)\right)$

Algorithme 2 : Calcul des variances de $v_p\left(\hat{\theta}^{\mathrm{R}}(t)\right)$ et $v_p\left(\hat{\theta}^{\mathrm{R}}(t) - \hat{\theta}(t)\right)$ par bootstrap généralisé.

Construction des estimateurs non robustes dans les réplications

Les expressions des estimateurs non robustes dépendent de l'approche choisie. Ainsi, pour l'estimation instant par instant et la troncature fonctionnelle, pour la réplication $b = 1, \ldots, \mathrm{B}$, on a

$$\hat{\theta}^{*(b)} = \sum_{i \in s} d_i \mathrm{W}_i^{*(b)} \mathrm{Y}_i. \tag{3.65}$$

En revanche, pour l'estimation par projection sur la base des composantes principales sphériques, cet estimateur est donné par

$$\hat{\theta}^{*(b)} = \mathrm{N}\hat{m}^{*(b)}(t) + \sum_{k=1}^{\mathrm{K}} \sum_{i \in s} d_i \mathrm{W}_i^{*(b)} f_{ik}^{*(b)} \hat{\mathrm{Z}}_k^{*(b)} \tag{3.66}$$

avec $\hat{m}^{*(b)}$ la solution de l'équation estimante non linéaire pondérée

$$\sum_{i \in s} d_i \mathrm{W}_i^{*(b)} \frac{\mathrm{Y}_i - \hat{m}^{*(b)}}{||\mathrm{Y}_i - \hat{m}^{*(b)}||} = 0, \tag{3.67}$$

$\hat{\mathrm{Z}}_k^{*(b)}$ les valeurs propres de la matrice de covariance $\hat{\Gamma}^{*(b)}$,

$$\hat{\Gamma}^{*(b)}(r,t) = \frac{1}{\mathrm{N}} \sum_{i \in s} \mathrm{W}_i^{*(b)} d_i \frac{(\mathrm{Y}_i(r) - \hat{m}^{*(b)}(r))}{||\mathrm{Y}_i - \hat{m}^{*(b)}||} . \frac{(\mathrm{Y}_i(t) - \hat{m}^{*(b)}(t))}{||\mathrm{Y}_i - \hat{m}^{*(b)}||}, \quad r, t \in [0, \mathrm{T}] \tag{3.68}$$

et $f_{ik}^{*(b)} = < \mathrm{Y}_i, \hat{\mathrm{Z}}_k^{*(b)} >$.

Estimation des biais conditionnels dans les réplications

Les biais conditionnels $\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}$ doivent être réestimés dans chaque rééchantillon. Ainsi, pour un estimateur de Horvitz-Thompson et un plan de sondage stratifié, on aura, en modifiant l'équation (3.7),

$$\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t) = \frac{\mathrm{N}_h}{\mathrm{N}_h - 1} \left(\frac{\mathrm{N}_h}{n_h} - 1)(\mathrm{Y}_i - \hat{\bar{\mathrm{Y}}}_h^{*(b)}) \right), \quad i \in s_h, \quad b = 1, \ldots, \mathrm{B} \tag{3.69}$$

avec

$$\hat{\bar{\mathrm{Y}}}_h^{*(b)} = \frac{\sum_{i\in s_h} \mathrm{W}_i^{*(b)}\mathrm{Y}_i}{\sum_{i\in s_h} \mathrm{W}_i^{*(b)}}, \quad h = 1,\dots \mathrm{H}, \quad b = 1,\dots \mathrm{B}.$$

Construction des estimateurs robustes dans les réplications

L'expression des estimateurs robustes dépendra de l'approche utilisée ainsi que du critère de détermination de la constante. Ainsi, pour l'approche instant par instant et notre nouveau critère défini en (3.45) pour une puissance q entière, pour la réplication $b = 1,\dots,\mathrm{B}$, on a

$$\hat{\theta}^{*\mathrm{R}(b)}(t) = \sum_{i\in s}\frac{1}{\pi_i}\mathrm{W}_i^{*(b)}\mathrm{Y}_i(t) + \sum_{i\in s}\mathrm{W}_i^{*(b)}\psi_{c_{opt}^{*(b)}(t)}\left(\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t)\right) - \sum_{i\in s}\mathrm{W}_i^{*(b)}\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t), \quad t\in[0,\mathrm{T}], \tag{3.70}$$

avec les constantes de troncature $c_{opt}^{*(b)}(t), \quad t\in[0,\mathrm{T}]$ déterminées en résolvant le problème d'optimisation

$$c_{opt}^{*(b)}(t) = \arg\min_{c\geq 0}\sum_{i\in s}\left|\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t) - \Delta^{*(b)}(c)\right|^q, \quad t\in[0,\mathrm{T}]$$

où

$$\Delta^{*(b)}(c) = \sum_{i\in s}\mathrm{W}_i^*\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t) - \sum_{i\in s}\mathrm{W}_i^*\psi_c\left(\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t)\right).$$

Pour l'approche instant par instant et le critère minimax, il n'est pas nécessaire d'estimer explicitement la constante. On préconise alors de remplacer le minimum et le maximum par les quantiles à 1% et à 99% dans l'expression de l'estimateur robuste univarié. On a alors

$$\hat{\theta}^{*\mathrm{R}(b)}(t) = \hat{\theta}^{*(b)}(t) - \frac{1}{2}\left(\hat{\mathrm{B}}_{q1}^{\hat{\theta}*(b)}(t) + \hat{\mathrm{B}}_{q99}^{\hat{\theta}*(b)}(t)\right), \quad t\in[0,\mathrm{T}], \tag{3.71}$$

avec $\hat{\mathrm{B}}_{q1}^{\hat{\theta}*(b)}(t)$ et $\hat{\mathrm{B}}_{q99}^{\hat{\theta}*(b)}(t)$ respectivement les quantiles à 1% et à 99% des biais conditionnels estimés pour la réplication b et l'estimateur $\hat{\theta}$.

Pour l'estimation par projection dans la base de l'ACP, on a

$$\hat{\theta}^{*\mathrm{R}(b)} = \mathrm{N}\hat{m}^{*(b)}(t) + \sum_{k=1}^{\mathrm{K}}\left(\sum_{i\in s}\frac{1}{\pi_i}\mathrm{W}_i^{*(b)}f_{ik}^{*(b)} + \mathrm{W}_i^{*(b)}\psi_{c_{opt}^{*(b)}(k)}(\hat{\mathrm{B}}_{1i}^{\hat{\mathrm{F}}_k,*(b)}) - \mathrm{W}_i^{*(b)}\hat{\mathrm{B}}_{1i}^{\hat{\mathrm{F}}_k,*(b)}\right)\hat{\mathrm{Z}}_k^{*(b)}(t) \tag{3.72}$$

avec $c_{opt}^{*(b)}(k)$ la constante optimale de troncature pour la composante k déterminée sur le rééchantillon b (d'une façon similaire à ce qui a été fait dans l'approche instant par instant), et $\hat{\mathrm{B}}_{1i}^{\hat{\mathrm{F}}_k,*(b)}$ le biais conditionnel estimé de l'unité i pour l'estimateur non robuste du total des scores de la composante k sur le rééchantillon b.

Enfin pour la troncature fonctionnelle, l'estimateur robuste pour la réplication b est

$$\hat{\theta}^{*\mathrm{R}(b)}(t) = \sum_{i\in s} \frac{1}{\pi_i} \mathrm{W}_i^{*(b)} \mathrm{Y}_i(t) + \sum_{i\in s} \mathrm{W}_i^{*(b)} \psi_{\alpha_{opt}^{*(b)}} \left(\hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}\right)(t) - \sum_{i\in s} \mathrm{W}_i^{*(b)} \hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t), \quad t \in [0,\mathrm{T}], \tag{3.73}$$

avec $\alpha_{opt}^{*(b)}$ la constante optimale de troncature fonctionnelle déterminée sur le rééchantillon b. Pour notre nouveau critère défini dans (3.47), cette constante est la solution du problème de minimisation

$$\alpha_{opt}^{*(b)} = \arg\min_{\alpha\geq 0} \left(\sum_{i\in s} \sum_{l=1}^{\mathrm{L}} \left| \hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t_l) - \Delta_\alpha^{*(b)}(t_l) \right|^q \right), \tag{3.74}$$

où

$$\Delta_\alpha^{*(b)}(t) = \sum_{j\in s} \mathrm{W}_j^{*(b)} \hat{\mathrm{B}}_{1j}^{\hat{\theta},*(b)}(t) - \mathrm{W}_j^{*(b)} \psi_\alpha \left(\hat{\mathrm{B}}_{1j}^{\hat{\theta},*(b)}\right)(t), \tag{3.75}$$

avec ψ_α défini par l'équation (3.40) mais en recalculant les profondeurs, la zone centrale ainsi que les bornes supérieures et inférieures sur le rééchantillon.

De même, pour le critère minimax, on adapte le critère :

$$\alpha_{opt}^{*(b)} = \arg\min_{\alpha\geq 0} \left(\max_{i\in s} \sum_{l=1}^{\mathrm{L}} \left| \hat{\mathrm{B}}_{1i}^{\hat{\theta},*(b)}(t_l) - \Delta_\alpha^{*(b)} \right| \right) \tag{3.76}$$

avec $\Delta_\alpha^{*(b)}$ défini par (3.75).

3.6 Application sur des jeux de données réelles

Dans cette section, nous testons et comparons entre eux les différents estimateurs proposés pour l'estimation robuste de courbe moyenne mais aussi l'estimation d'Erreur Quadratique Moyenne (EQM, ou MSE pour *Mean Squared Error*) sur un jeu de données réel de consommation électrique. On s'attache en particulier à quantifier le gain de précision induit par l'estimation robuste par rapport à l'estimation non robuste pour différentes stratégies d'estimation (i.e. combinaison d'un plan de sondage et d'un estimateur).

Après avoir présenté quelques statistiques descriptives sur le jeu de données qui témoignent du risque de présence d'unités influentes dans les échantillons, on propose des tests basés sur le plan de sondage, d'une part pour évaluer la qualité de nos estimateurs de courbes totales dans la sous-section 3.6.2 puis d'autre part pour évaluer la qualité des estimateurs d'erreur quadratique moyenne dans la sous-section 3.6.3.

3.6.1 Présentation du jeu de données

On travaille sur des données en accès libre issues d'une expérimentation de la commission irlandaise pour la régulation de l'énergie (CER). Ces données ont été collectées dans le cadre d'un projet sur les compteurs communicants qui s'est déroulé en 2009-2010 (CER, 2011). Le jeu de données d'origine est disponible sur demande à l'adresse `http://www.ucd.ie/issda/data/commissionforenergyregulation/`.

Parmi ces courbes, nous avons sélectionné N = 3994 courbes de clients résidentiels au pas demi horaire pour la semaine du 18 au 24 janvier 2010 (soit L = 336 points demi-horaires). Ces données ne comportent pas de valeurs manquantes pour la période considérée. Nous avons également calculé, pour chaque ménage, sa consommation moyenne sur le semestre précédent (juin à décembre 2009). Cette information à une maille temporelle très agrégée est utilisée au niveau du plan de sondage dans les plans stratifiés. Par ailleurs, la consommation moyenne sur la semaine d'étude est quant à elle utilisée comme variable de calage : en effet, cette information ne peut par définition être connue qu'a posteriori.

Les N = 3994 ménages considérés constituent notre population d'intérêt. De premiers exemples de courbes de consommation électrique ont déjà été présentés dans la section 2.1.1 et ont permis d'illustrer la grande irrégularité des courbes ainsi que leurs saisonnalité.

Quelques statistiques descriptives La distribution des niveaux moyens des courbes est extrêmement asymétrique, comme l'atteste l'histogramme des moyennes des courbes sur la semaine d'étude de la Figure 3.6. Cette asymétrie est marquée également si l'on s'intéresse aux courbes des quantiles instantanés des consommations (Figure (3.7)). Nos échantillons peuvent donc contenir des unités atypiques dont le niveau est très sensiblement supérieur à celui des autres unités du panel.

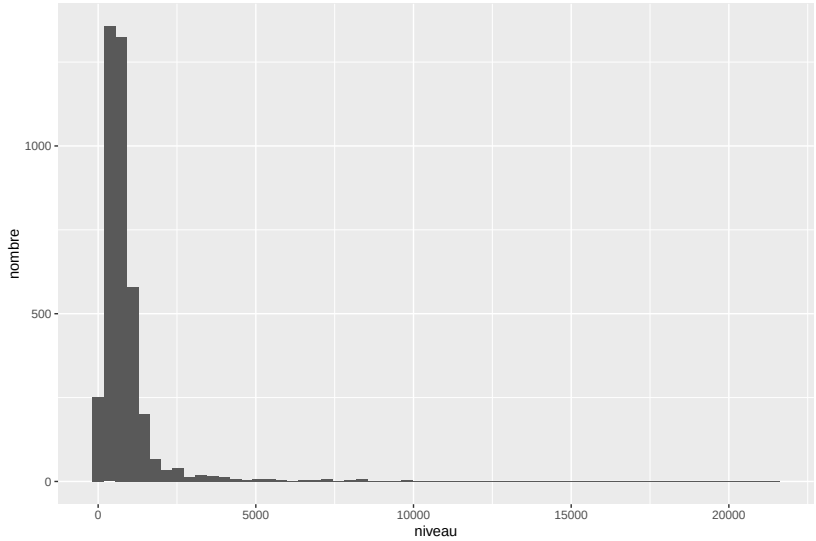

FIGURE 3.6 – Histogrammes des puissances moyennes en W sur la semaine d'étude

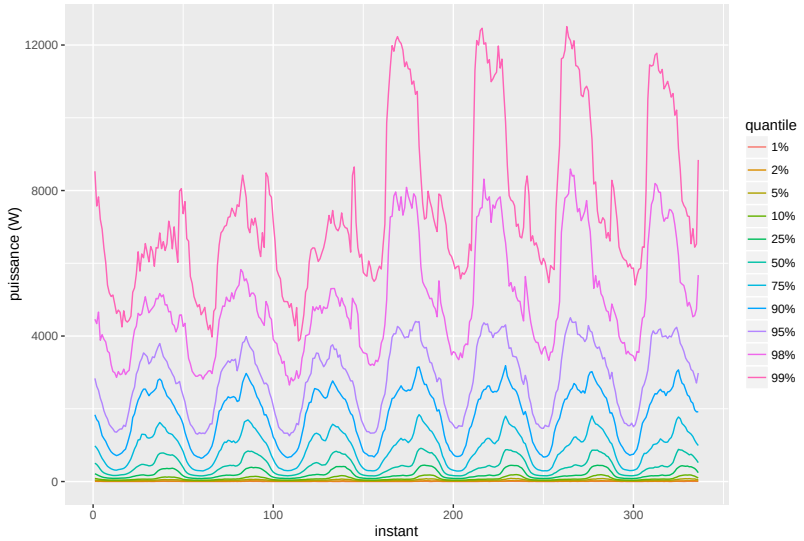

FIGURE 3.7 – Quantiles instantanés des consommations

3.6.2 Tests des estimateurs robustes de courbes moyennes

Dans cette sous-section, nous comparons les performances de nos différents estimateurs robustes de courbe moyenne, entre eux et avec l'estimateur non robuste associé, pour différentes stratégies d'estimation (*i.e.* plan de sondage et estimateur non robuste) et différentes tailles d'échantillon. En particulier, les strata jumpers, c'est-à-dire les unités affectées à de mauvaises strates durant la constitution du plan de sondage du fait d'information auxiliaire erronée, peuvent constituer des unités influentes et on simulera donc la présence de telles unités dans nos tests afin d'évaluer la capacité de nos estimateurs à limiter leur impact. L'ensemble des tests ont été réalisés en R. Pour implémenter l'estimateur par calage, nous avons utilisé la fonction `calib` du package `sampling`.

Protocole de test

Nous comparons les estimateurs suivants :

- Estimateur non robuste de Horvitz-Thompson ou par calage selon la stratégie d'estimation (notés *non robuste*)
- Estimateur non par décomposition dans une base d'ondelettes [4] sans seuillage (*non robuste ondelettes*) ou avec seuillage (*non robuste ondelettes seuillées*)
- Estimateur robuste construit par troncature univariée par instant de discrétisation décrit au paragraphe 3.4.1, avec choix de la constante de troncature par le critère minimax (*ponctuel minimax*)
- Estimateur robuste construit par troncature univariée par instant de discrétisation décrit au paragraphe 3.4.1, avec choix de la constante de troncature par le critère de la somme des biais conditionnels, pour la constante $q = 4, 10, 20$ défini dans 3.4.4 (*robuste sumbias q*)

4. ondelettes de Daubechies Least Asymetric, 10

- Estimateur robuste construit par troncature univariée sur les composantes principales de l'ACP sphérique décrit dans le paragraphe 3.4.2, critère minimax (*robuste acp*)
- Estimateur robuste construit par troncature univariée sur les coordonnées des courbes dans une base d'ondelettes [5] sans seuillage (*ponctuel ondelettes*).
- Estimateur robuste par troncature fonctionnelle présenté dans le paragraphe 3.4.3, basé sur la profondeur MBD, critère minimax (*fonctionnel mbd*)
- Estimateur robuste par troncature fonctionnelle basée sur la profondeur déduite de la projection sur les 5 premières composantes principales sphériques, critère minimax (*fonctionnel projection acp*)
- Estimateur robuste par troncature fonctionnelle basée sur la décomposition en ondelettes, avec seuillage dur, critère minimax (*fonctionnel projection ondelettes*)

Nous nous intéressons aux scénarios d'estimation suivants (combinaison d'un estimateur, d'un plan de sondage et d'un taux de strata jumpers le cas échéant) :

- SAS : Sondage aléatoire simple et estimateur de Horvitz-Thompson
- STR SJ0 : Sondage aléatoire stratifié (sur la consommation des six mois précédents) avec allocation optimale de Neyman et estimateur de Horvitz-Thompson, sans strata jumper
- STR SJ10 : Sondage aléatoire stratifié (sur la consommation des six mois précédents) avec allocation optimale de Neyman et estimateur de Horvitz-Thompson, avec 10% de strata jumpers
- STR SJ20 : Sondage aléatoire stratifié (sur la consommation des six mois précédents) avec allocation optimale de Neyman et estimateur de Horvitz-Thompson, avec 20% de strata jumpers
- SAS calage : Sondage Aléatoire Simple et estimateur par calage (sur la consommation durant la semaine de test)

Pour le sondage stratifié, nous divisons la population en cinq strates de consommation homogène, dont les limites sont définies de façon à ce que la somme de la variable de stratification (consommation moyenne des six mois précédents) soit égale dans chacune des strates. On aura donc davantage d'individus dans les plus petites strates de consommation.

Notre protocole de test consiste à tirer aléatoirement un grand nombre d'échantillons dans notre population d'intérêt pour chaque plan de sondage et chaque taille d'échantillon puis à estimer la courbe moyenne de la population à partir de chacun de ces échantillons par les différents estimateurs robustes et non robustes. Ensuite, nous comparons les estimations obtenues à la courbe de consommation moyenne de la population, connue, pour en déduire des indicateurs de qualité.

Plus précisément, pour un plan de sondage, une taille d'échantillon, un taux de strata jumpers τ_{sj} et un nombre de simulations E donnés, le protocole de test est celui décrit dans l'algorithme 3. Pour chaque scénario de tests, nous avons considéré 4 tailles

5. ondelettes de Daubechies Least Asymetric, 10

d'échantillon différentes $n = 400, 200, 100, 40$. Dans chaque cas, nous tirons $\mathrm{E} = 5000$ échantillons.

Pour simuler des strata jumpers dans les plans stratifiés, nous procédons de la manière suivante : pour chaque simulation, avant l'étape de tirage de l'échantillon, nous sélectionnons aléatoirement τ_{sj} % des unités de la population par sondage aléatoire simple, et pour chaque unité nous tirons aléatoirement et avec probabilité uniforme une "mauvaise" strate parmi les strates auxquelles l'unité n'appartient pas puis nous remplaçons sa véritable strate par cette "mauvaise" strate.

Soit E le nombre de réplications, (E grand).
pour $i \leftarrow 1$ ***à*** E **faire**
- Simuler l'apparition des strata jumpers en tirant aléatoirement τ_{sj}% des unités de la population U et remplacer leur strate par une strate incorrecte choisie aléatoirement.
- Tirer un échantillon s à partir de cette population modifiée, selon le plan de sondage p, pour une taille d'échantillon n.
- Estimer la courbe moyenne de la population μ_{Y}^{e} par les différents estimateurs cités plus haut et stocker les estimations.

fin
A partir des estimations $\hat{\mu}_{\mathrm{Y}}^{e}, \quad e = 1, \ldots, \mathrm{E}$ obtenues à chaque itération, calculer les indicateurs de qualité RB, MSE, RE selon les formules (3.77) à (3.79).
retourner RB,MSE,RE

Algorithme 3 : Comparaison des performances des différents estimateurs de courbes moyennes, pour un plan de sondage p, une taille d'échantillon n, et un nombre de réplications E.

Indicateurs de qualité

On note $\mathbb{E}_{\mathrm{MC}}[\hat{\mu}_{\mathrm{Y}}(t_l)] = \frac{1}{\mathrm{E}} \sum_{e=1}^{\mathrm{E}} \hat{\mu}_{\mathrm{Y}}^{e}(t_l)$ "l'espérance de Monte-Carlo" de l'estimateur $\hat{\mu}_{\mathrm{Y}}$ pour l'instant t_l avec $\hat{\mu}_{\mathrm{Y}}^{e}$ l'estimateur de la courbe moyenne obtenue à la simulation e.

Pour un instant de discrétisation donné $t_l, \quad l = 1, \ldots, \mathrm{L}$, on construit les indicateurs de qualité suivants :

$$\mathrm{RB}(\hat{\mu}_{\mathrm{Y}})(t_l) = 100 \frac{\mathbb{E}_{\mathrm{MC}}[\hat{\mu}_{\mathrm{Y}}(t_l)] - \mu_{\mathrm{Y}}(t_l)}{\mu_{\mathrm{Y}}(t_l)}, \tag{3.77}$$

$$\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_{\mathrm{Y}})(t_l) = \frac{1}{\mathrm{E}} \sum_{e=1}^{\mathrm{E}} \left(\hat{\mu}_{\mathrm{Y}}^{e}(t_l) - \mu_{\mathrm{Y}}(t_l)\right)^2, \tag{3.78}$$

avec $\mu_{\mathrm{Y}}(t_l)$ la véritable valeur de la courbe moyenne de la population U pour cet instant. Le premier indicateur appelé Biais Relatif (noté RB pour *Relative Bias*) quantifie le biais sous le plan de sondage d'un estimateur, et le second, $\mathrm{MSE}_{\mathrm{MC}}$ quantifie l'erreur globale (carré du biais plus variance). Plus la valeur de ces indicateurs est basse (en valeur absolue pour le RB), plus l'estimateur est considéré comme performant.

L'erreur quadratique moyenne peut être difficile à interpréter, on va donc utiliser un troisième indicateur plus facile à lire appelé Efficacité Relative (RE pour *Relative Efficiency*), qui compare l'erreur quadratique moyenne Monte-Carlo $\mathrm{MSE}_{\mathrm{MC}}$ de chaque méthode avec celle d'un estimateur de référence. Ici, pour chaque configuration (estimateur non robuste et plan de sondage), on cherche à comparer les performances des estimateurs robustes à celles de l'estimateur non robuste $\hat{\mu}_{\mathrm{Y}}^{0} = \hat{\mu}_{\mathrm{Y}}^{\mathrm{NROB}}$, on utilisera donc l'estimateur normalisé suivant :

$$\mathrm{RE}(\hat{\mu}_{\mathrm{Y}})(t_l) = 100\frac{\mathrm{MSE}_{\mathrm{MC}}[\hat{\mu}_{\mathrm{Y}}(t_l)]}{\mathrm{MSE}_{\mathrm{MC}}[\hat{\mu}_{\mathrm{Y}}^{0}(t_l)]}. \tag{3.79}$$

Plus l'indicateur RE sera faible, plus l'estimateur sera considéré comme performant. Un RE de 100 correspond à un indicateur aussi performant que l'estimateur non robuste.

Afin d'évaluer la performance globale, on considère finalement la moyenne de ces indicateurs sur l'ensemble des instants de la période de test :

$$\mathrm{RB}(\hat{\mu}_{\mathrm{Y}}) = \frac{1}{\mathrm{L}}\sum_{l=1}^{\mathrm{L}} \mathrm{RB}(\hat{\mu}_{\mathrm{Y}})(t_l), \tag{3.80}$$

$$\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_{\mathrm{Y}}) = \frac{1}{\mathrm{L}}\sum_{l=1}^{\mathrm{L}} \mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_{\mathrm{Y}})(t_l), \tag{3.81}$$

$$\mathrm{RE}(\hat{\mu}_{\mathrm{Y}}) = \frac{1}{\mathrm{L}}\sum_{l=1}^{\mathrm{L}} \mathrm{RE}(\hat{\mu}_{\mathrm{Y}})(t_l). \tag{3.82}$$

On s'intéresse également aux temps de calcul des différents estimateurs.

Résultats pour les estimateurs de courbe moyenne

Les performances des différentes approches sont résumées dans les Tables 3.1 à 3.6 et les Figures 3.8 à 3.16.

Les méthodes robustes induisent d'importants gains de précision, en particulier lorsque la taille de l'échantillon est faible. Ainsi, pour un sondage aléatoire simple et la meilleure méthode robuste, l'erreur globale est réduite de 28 % pour une taille d'échantillon de 40, de 17% pour une taille d'échantillon de 100 et de 10% pour une taille d'échantillon de 200 (voir Table 3.2). De plus, les méthodes robustes ne dégradent jamais significativement la précision globale. On remarque cependant que, pour l'estimateur par calage ou pour le plan de sondage stratifié sans strata jumpers, les méthodes robustes par troncature fonctionnelle peuvent conduire à des résultats proches de ceux des méthodes non robustes voire légèrement moins bons, ce qui n'est jamais le cas des méthodes robustes par troncature univariée.

Les performances des trois approches sont assez proches mais on constate quand même un léger avantage pour la troncature univariée sur les scores de l'ACP robustes ou les cœfficients d'ondelettes (voir Figure 3.10). Les méthodes fonctionnelles par projection et la troncature instant par instant arrivent ensuite, suivie par la troncature

fonctionnelle selon la Modified Band Depth. Globalement, la projection sur les bases d'ondelettes donne des résultats légèrement meilleurs que la projection sur la base de l'ACP robuste.

Cependant, les méthodes robustes ont tendance à sous-estimer la courbe moyenne réelle, ce qui est attendu car dans notre cas d'application, du fait de l'asymétrie de la distribution des consommations moyennes, les unités influentes dont nous cherchons à borner l'impact sont justement des unités dont les valeurs sont élevées. Les méthodes robustes induisent donc un biais relatif de quelques pour-cents, qui est d'autant plus grand que les plans de sondage sont imprécis et les tailles d'échantillons petites (voir Figure 3.8).

Pour les sondages stratifiés sans strata jumper, les méthodes robustes engendrent de plus petits gains de précision que pour les sondages aléatoires simples : en effet la stratification permet de donner de plus petits poids de sondage aux plus grosses unités grâce à l'utilisation de l'allocation optimale de Neyman pour la détermination des tailles de strates. L'influence des grosses unités est donc déjà réduite au cours de la phase d'échantillonnage et les estimateurs robustes permettent des améliorations plus faibles (entre 5 et 10%), la meilleure méthode étant en général l'estimation robuste par troncature dans la base des ondelettes. On remarque que dans ce cas le biais relatif est faible (moins de 3%) ce qui pourrait vouloir dire que les biais conditionnels ne sont que peu tronqués et l'estimateur robuste peu différent de l'estimateur non robuste.

Au contraire, en présence de strata jumpers, les méthodes robustes limitent la perte de précision par rapport à la situation sans strata jumper surtout lorsque leur taux est élevé. Les gains de précision par rapport à l'estimateur robuste sont alors autour de 15% pour un taux de strata jumpers de 10%.

Pour un estimateur par calage et un plan de sondage aléatoire simple, les gains de précision induits par la robustesse sont sensiblement plus faibles : entre 6 et 8% pour la meilleure méthode (troncature dans las base des ondelettes), pour l'ensemble des tailles d'échantillon.

Le critère minimax induit de meilleurs résultats que notre nouveau critère. Cela se vérifie pour les trois approches étudiées. Néanmoins, comme attendu, à mesure que q augmente, les performances des deux critères deviennent comparables (voir Figure 3.14). Notre nouveau critère semble dégrader les performances des estimateurs, on privilégiera donc le minimax.

Parmi les méthodes robustes, la troncature instant par instant est la plus rapide avec un temps de calcul pour une simulation autour de six centièmes de secondes pour l'estimateur de Horvitz-Thompson et un échantillon de 100 individus. Les méthodes par projection sont légèrement plus lentes mais toujours en dessous de deux dixièmes de secondes et enfin les troncatures fonctionnelles sont les plus coûteuses en temps de calcul, autour de 3 dixièmes de secondes pour des échantillons de 100 individus. Le calage augmente les temps de calcul, pour les méthodes robustes ou non. Globalement ces temps d'exécution restent toutefois très raisonnables (voir Figure 3.16).

L'erreur quadratique moyenne est très variable au cours du temps : de manière logique, elle est plus importante lorsque la valeur de la courbe moyenne est la plus élevée. Inversement, les gains de précision sont plus importants là où on était auparavant le plus imprécis, c'est pourquoi l'Efficacité Relative est plus faible à ces mêmes instants. L'interclassement relatif des erreurs quadratiques moyennes de Monte-Carlo ne

varie cependant pas au cours du temps, comme l'atteste la Figure 3.15.

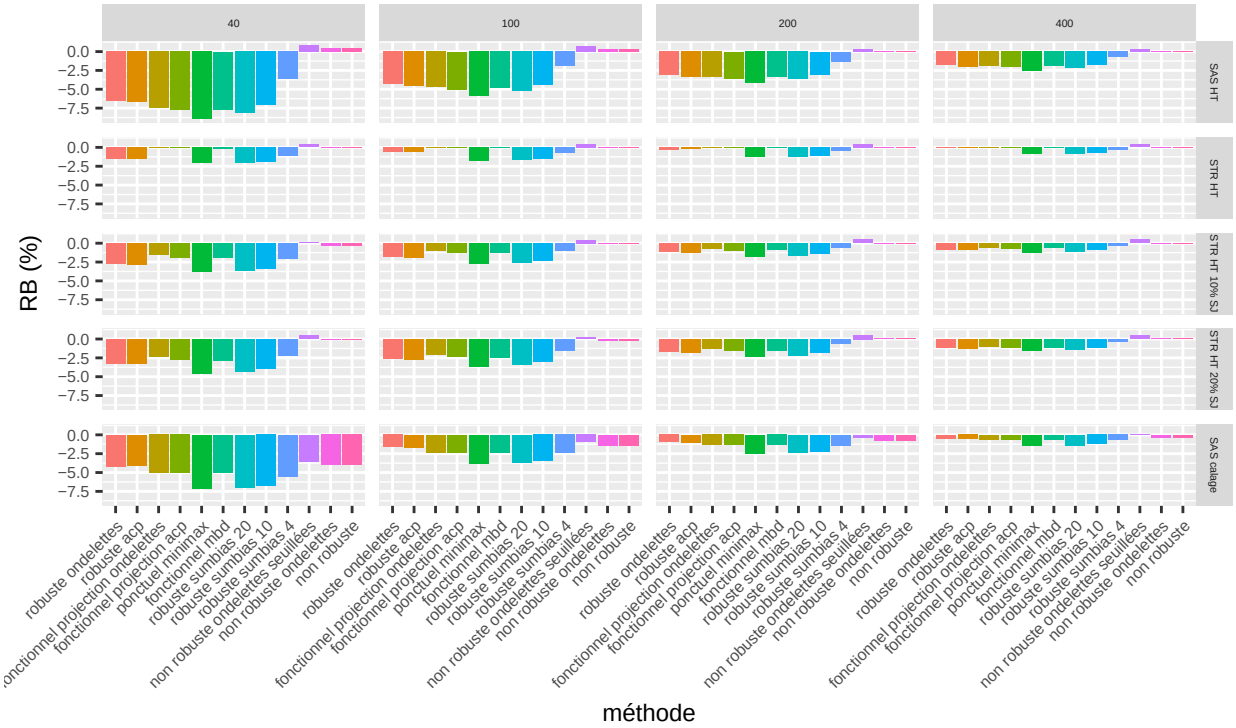

FIGURE 3.8 – Biais relatifs (RB en %), voir (3.80)) des différents estimateurs pour les différents plans de sondage en fonction des tailles d'échantillon

taille	SAS HT	STR HT	STR HT 10% SJ	STR HT 20% SJ	SAS calage
40	63465	25392	30896	38731	30352
100	24853	9767	12204	14742	12404
200	11912	4703	5843	7127	6107
400	5746	2115	2663	3351	2856

TABLEAU 3.1 – MSE de l'estimateur non robuste par taille d'échantillon (taille) pour les différentes stratégies d'estimation

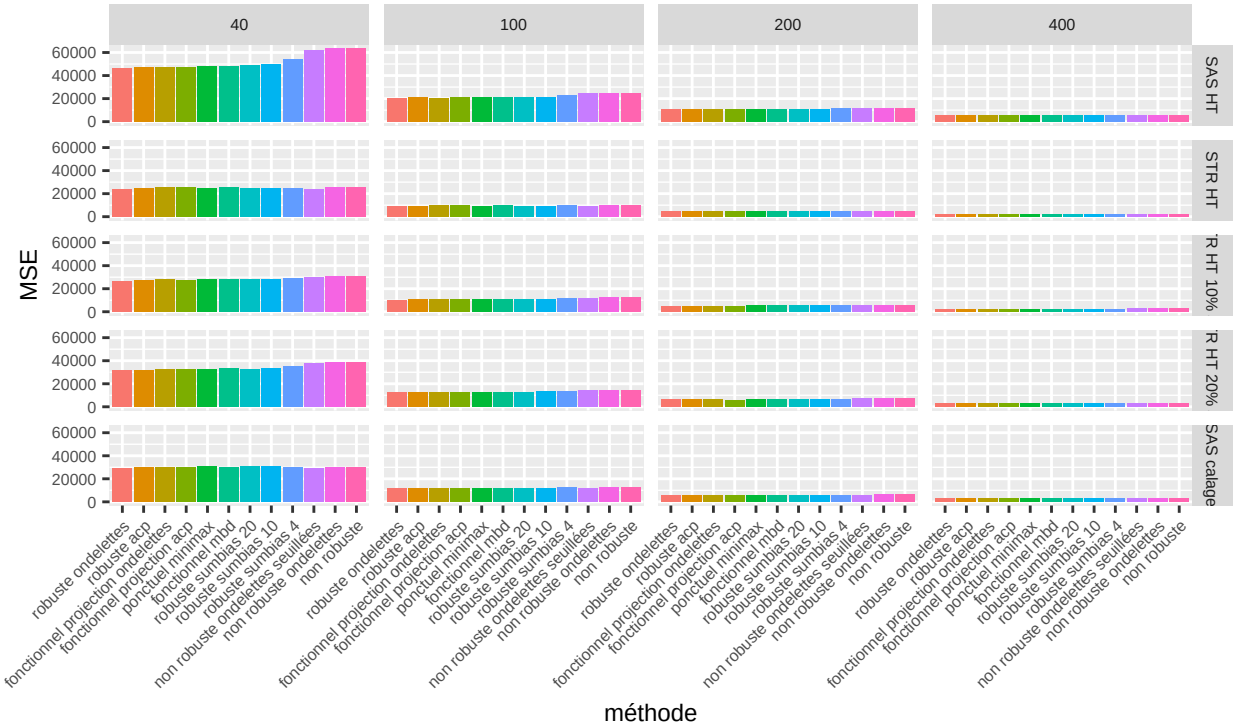

FIGURE 3.9 – Erreurs quadratiques moyennes (MSE), voir (3.81)) des différents estimateurs pour les différents plans de sondage en fonction des tailles d'échantillon

3.6.3 Tests des estimateurs d'erreur quadratique moyenne

Nous allons maintenant évaluer les performances des différents estimateurs d'Erreur Quadratique Moyenne exposés dans la section 3.5. On réalise des tests uniquement pour trois des scénarios de tests présentés plus haut : sondage aléatoire simple et estimateur de Horvitz-Thompson (SAS HT), sondage aléatoire simple par calage (SAS Calage) et enfin estimateur de Horvitz-Thompson et sondage stratifié sans strata jumper (STR HT), et pour une taille d'échantillon de 100 individus dans tous les cas.

Nous étudions les performances des estimateurs d'erreur quadratique moyenne pour l'estimateur non robuste (Horvitz-Thompson ou calage) mais aussi pour les versions robustes des estimateurs résultant de chacune des approches proposées, et pour les deux critères abordés : le critère minimax et notre nouveau critère présenté au paragraphe 3.4.4. Dans les différents cas, on compare les résultats obtenus par l'application des estimateurs de MSE explicites, par bootstrap populationnel et par bootstrap généralisé. Afin de ne pas multiplier à l'excès les scénarios de tests, on ne testera pas toutes les combinaisons possibles des méthodes d'estimation robustes × critère de troncature × méthodes d'estimation d'erreur quadratique moyenne mais seulement quelques scénarios de tests qui permettent de couvrir les principaux cas de figure qui nous intéressent :

- Estimateur de MSE explicite pour l'estimateur non robuste de courbe moyenne (noté *explicite non robuste*), équation 2.10
- Estimateur de MSE par bootstrap généralisé pour l'estimateur non robuste de

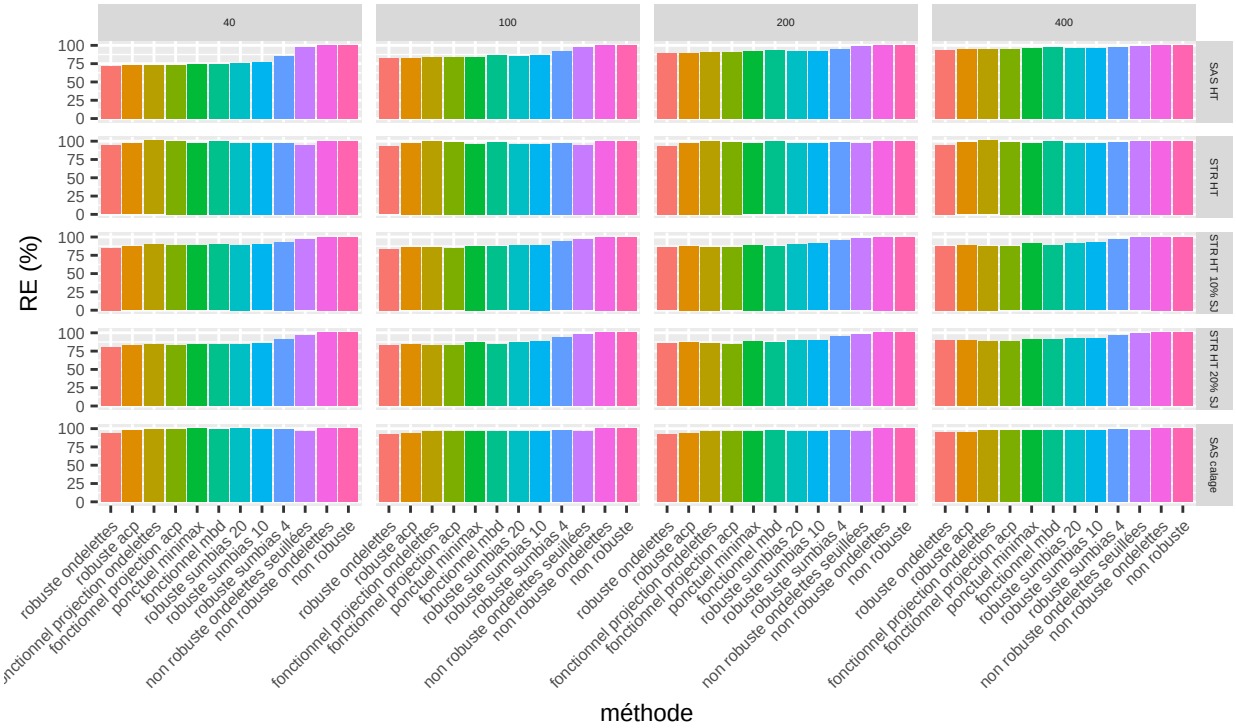

FIGURE 3.10 – Efficacités relatives (RE en %), voir (3.82)) des différents estimateurs pour les différents plans de sondage en fonction des tailles d'échantillon

courbe moyenne (*généralise non robuste*), équation 2.35

- Estimateur de MSE par bootstrap populationnel pour l'estimateur non robuste de courbe moyenne (*populationnel non robuste*), équation 2.35
- Estimateur de MSE explicite pour l'estimateur non robuste de courbe moyenne avec projection sur la base de l'ACP (*explicite non robuste acp*)
- Estimateur de MSE explicite pour l'estimateur non robuste de courbe moyenne avec projection sur une base d'ondelettes (*explicite non robuste ondelettes*)
- Estimateur explicite de MSE pour l'estimateur robuste par troncature univariée par instant de discrétisation, avec choix de la constante de troncature par le critère minimax (*explicite robuste minimax*), équation 3.51
- Estimateur de MSE par bootstrap généralisé pour l'estimateur robuste par troncature univariée par instant de discrétisation, avec choix de la constante de troncature par le critère minimax (*généralisé robuste minimax*) , voir paragraphe 3.5.3
- Estimateur de MSE par bootstrap populationnel pour l'estimateur robuste par troncature univariée par instant de discrétisation, avec choix de la constante de troncature par le critère minimax (*populationnel robuste minimax*), voir paragraphe 3.5.2
- Estimateur explicite de MSE pour l'estimateur robuste par troncature univariée sur les composantes principales de l'ACP , avec choix de la constante de troncature par le critère minimax (*explicite robuste minimax*) , équation 3.60

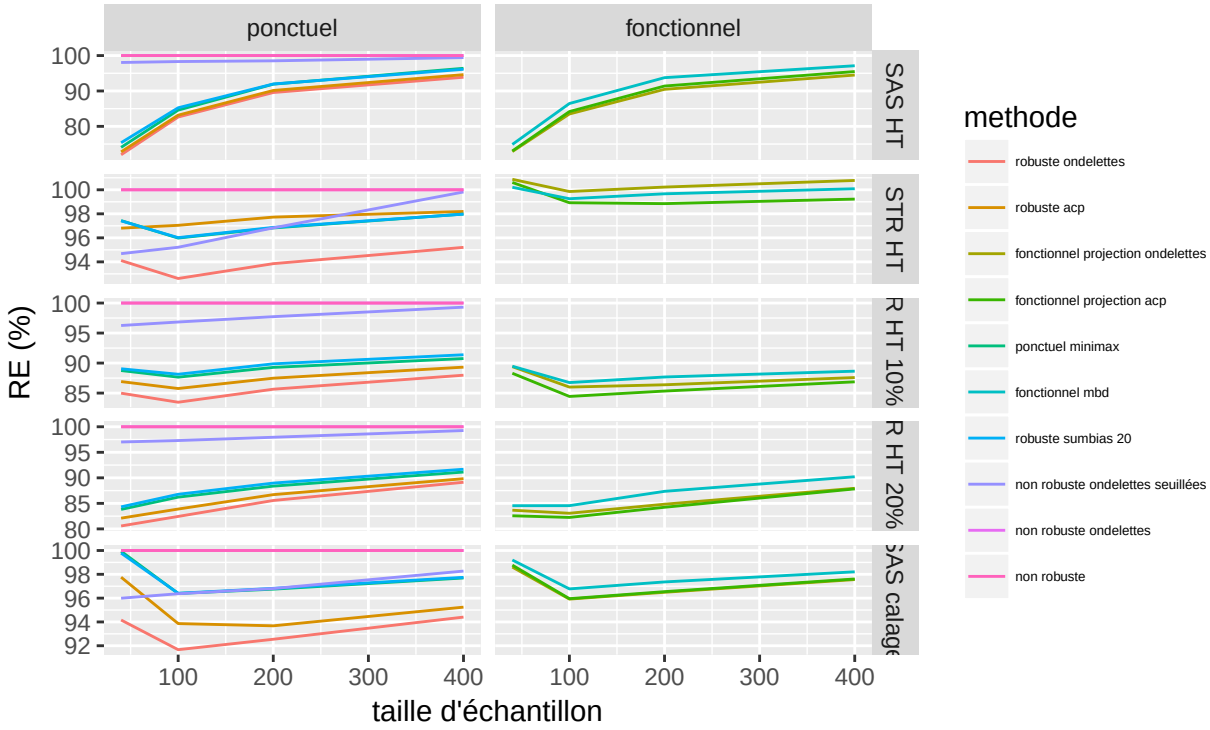

FIGURE 3.11 – Efficacités relatives (RE en %), voir (3.82)) des différents estimateurs pour le sondage aléatoire simple et en fonction des tailles d'échantillon

- Estimateur de MSE par bootstrap généralisé pour l'estimateur robuste par troncature univariée par instant de discrétisation, avec choix de la constante de troncature par le critère minimax (*généralisé robuste minimax*) , voir paragraphe 3.5.3
- Estimateur explicite de MSE pour l'estimateur robuste par troncature univariée par instant de discrétisation, avec choix de la constante de troncature par le critère de la somme des biais conditionnels à la puissance $q = 10$ (*explicite robuste sumbias*)
- Estimateur de MSE par bootstrap généralisé pour l'estimateur robuste par troncature univariée par instant de discrétisation, avec choix de la constante de troncature par le critère de la somme des biais conditionnels à la puissance $q = 10$ (*généralisé robuste sumbias*) , voir paragraphe 3.5.3

Procédure de test

La procédure de test utilisée ici est similaire à celle mise en œuvre pour évaluer la qualité des méthodes d'estimation : on tire un grand nombre d'échantillons dans notre population d'intérêt, puis sur chacun des échantillons on estime la courbe moyenne par les différents estimateurs ainsi que l'erreur quadratique moyenne associée (par les différentes méthodes d'estimation proposées). On compare finalement les approximations d'erreur quadratique moyenne à l'erreur quadratique moyenne de Monte-Carlo (équation (3.77)) déduite des simulations. Nous avons réalisé E = 5000 simulations. Nous réalisons B = 1000 réplications dans les bootstraps et dans notre critère de choix

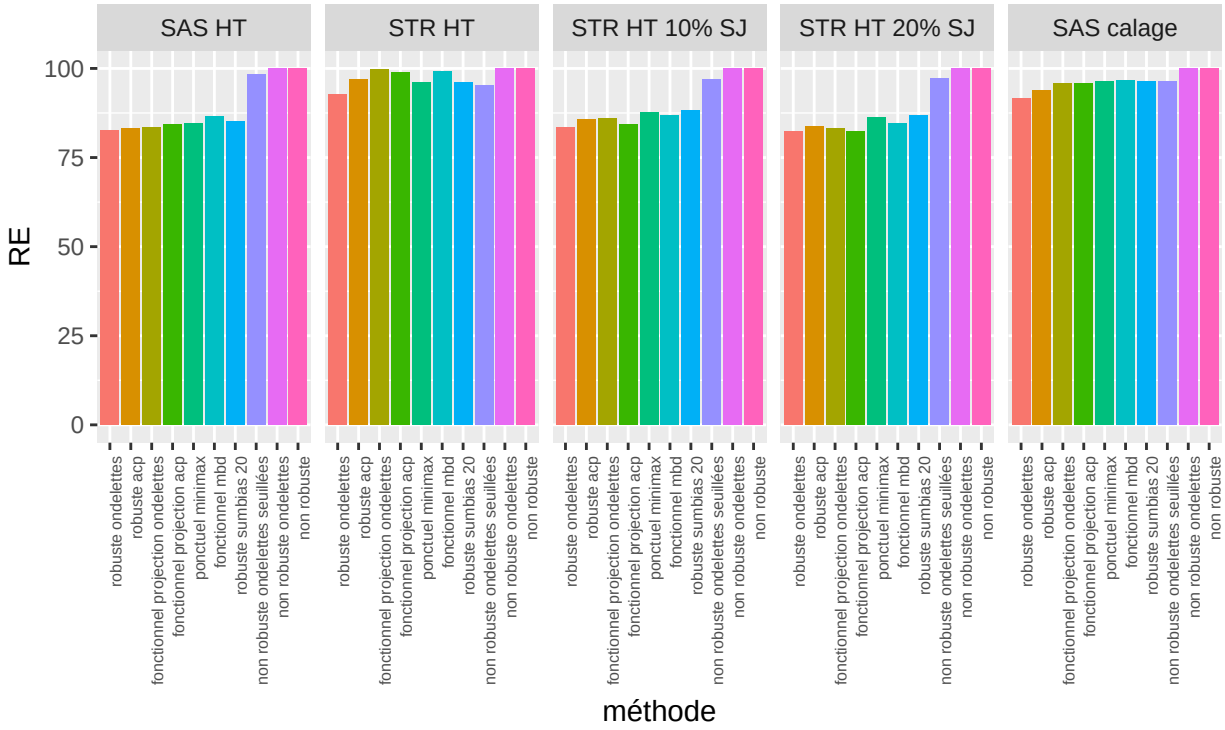

FIGURE 3.12 – Efficacités relatives (RE en %), voir (3.82)) des différents estimateurs et différents plans de sondages et une taille d'échantillon $n = 100$

de la constante, nous prenons $q = 10$. Pour les scénarios de test retenus, il n'est pas nécessaire de simuler des strata jumpers.

Indicateurs de qualité

Pour tout instant $t = t_1, \ldots, t_L$ On définit l'espérance Monte-Carlo de notre estimateur de MSE comme :

$$\mathbb{E}_{MC}[\widehat{MSE}(\hat{\mu}_Y)(t)] = \frac{1}{E}\sum_{e=1}^{E}\widehat{MSE}(\hat{\mu}_Y^e)(t). \tag{3.83}$$

Le critère de qualité considéré ici est le biais relatif de cet estimateur :

$$RB\left(\widehat{MSE}(\hat{\mu}_Y)(t)\right) = \frac{\mathbb{E}_{MC}[\widehat{MSE}(\hat{\mu}_Y)(t)] - MSE_{MC}\left(\hat{\mu}_Y\right)(t)}{MSE_{MC}\left(\hat{\mu}_Y\right)(t)}. \tag{3.84}$$

Plus ce biais relatif est faible en valeur absolue plus l'estimateur de MSE est considéré comme de bonne qualité. On s'intéresse également au temps de calcul pour l'ensemble des estimateurs.

Résultats pour l'estimation d'erreur quadratique moyenne

Les performances des différentes approximations d'erreur quadratique moyenne sont résumées dans les Tables 3.7 à 3.8 et illustrées par les Figures 3.17 à 3.19. La première conclusion de ces tests est que l'ensemble de ces estimateurs de MSE donnent

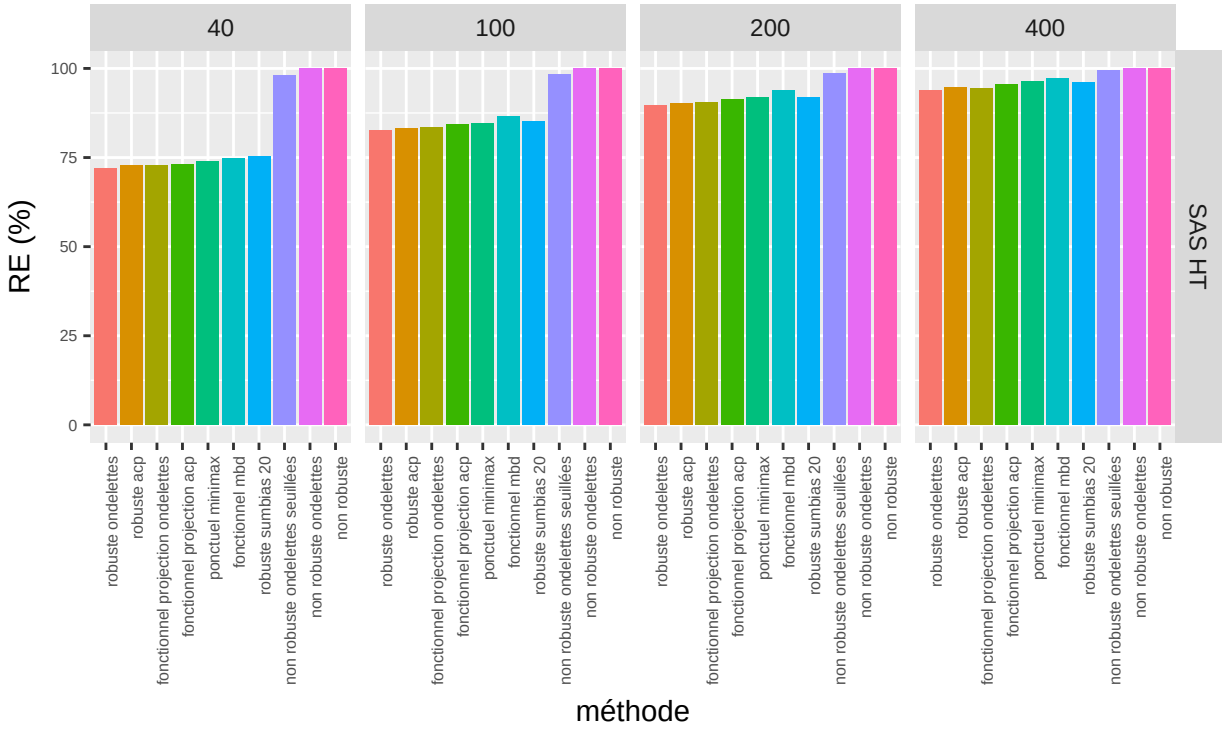

FIGURE 3.13 – Efficacités relatives (RE en %), voir (3.82)) des différents estimateurs pour le sondage aléatoire simple et différentes tailles d'échantillon

des résultats satisfaisants pour les estimateurs non robustes, sauf pour l'estimateur par calage pour lequel on constate une sous-estimation systématique de 30% pour la méthode explicite et 20% pour les bootstraps. Par ailleurs, pour le sondage aléatoire simple, pour les estimateurs robustes dont la constante est déterminée par le critère minimax ou notre nouveau critère, on constate une nette sous-estimation pour les estimateurs de MSE explicites : environ -20 % pour notre critère et -30% pour le minimax pour le sondage aléatoire simple. Cela est vraisemblablement dû au fait que l'on traite les constantes d'ajustement et donc les valeurs tronquées Y_i^* comme fixes alors qu'elles dépendent de l'échantillon. En revanche, lorsque l'on utilise un bootstrap avec la constante qui varie dans chaque rééchantillon, on constate a contrario une forte surestimation pour le bootstrap généralisé (30% pour le sondage aléatoire simple). En revanche, le bootstrap populationnel semble donner des résultats satisfaisants (surestimation de 3% pour le sondage aléatoire simple, sous-estimation de 6% pour le sondage stratifié, et l'estimateur robuste minimax). Pour l'estimateur par calage, on peut supposer que l'effet de sous-estimation observé sur l'estimateur non robuste s'additionne aux différents biais précédemment évoquées. On peut donc conclure que, **dans notre contexte particulier, le fait de considérer la constante comme fixe lorsqu'elle ne l'est pas induit une sous-estimation de l'erreur non négligeable**.

En ce qui concerne les temps de calcul, les méthodes explicites sont très rapides (entre un centième de seconde pour un estimateur de Horvitz-Thompson non robuste et 1.3 secondes pour l'estimateur par calage robuste).En revanche, les bootstraps sont beaucoup plus longs (80 secondes pour le bootstrap populationnel, 20 secondes pour

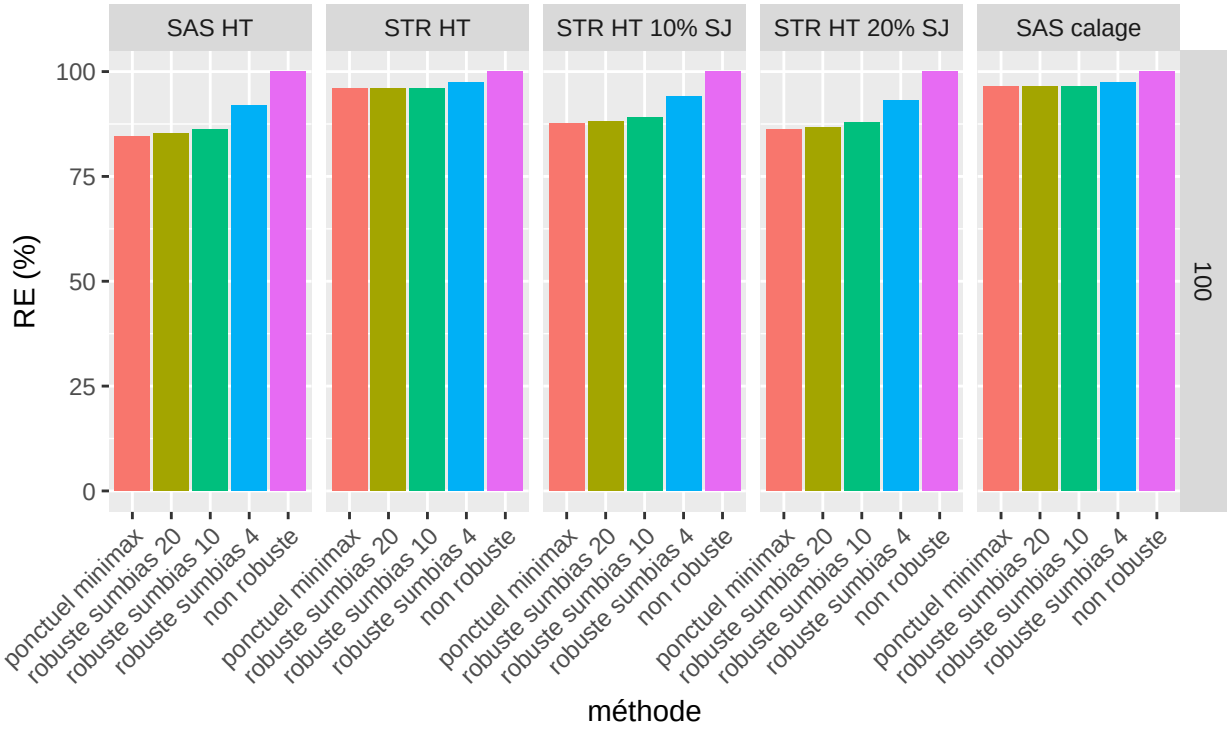

FIGURE 3.14 – Impact de la constante q dans le critère de la somme des biais conditionnels : RE des estimateurs robustes avec troncature univariée instant par instant dont la constante est choisie par le nouveau critère, en fonction de la puissance q, pour un échantillon de 100 unités

le bootstrap généralisé).

En conclusion si la taille des échantillons et le nombre d'instants considérés ne sont pas trop importants, on préfèrera donc les méthodes d'estimation par bootstrap populationnel : en effet, celui ci semble nettement plus proche de la réalité que les estimateurs explicites, certes plus rapides.

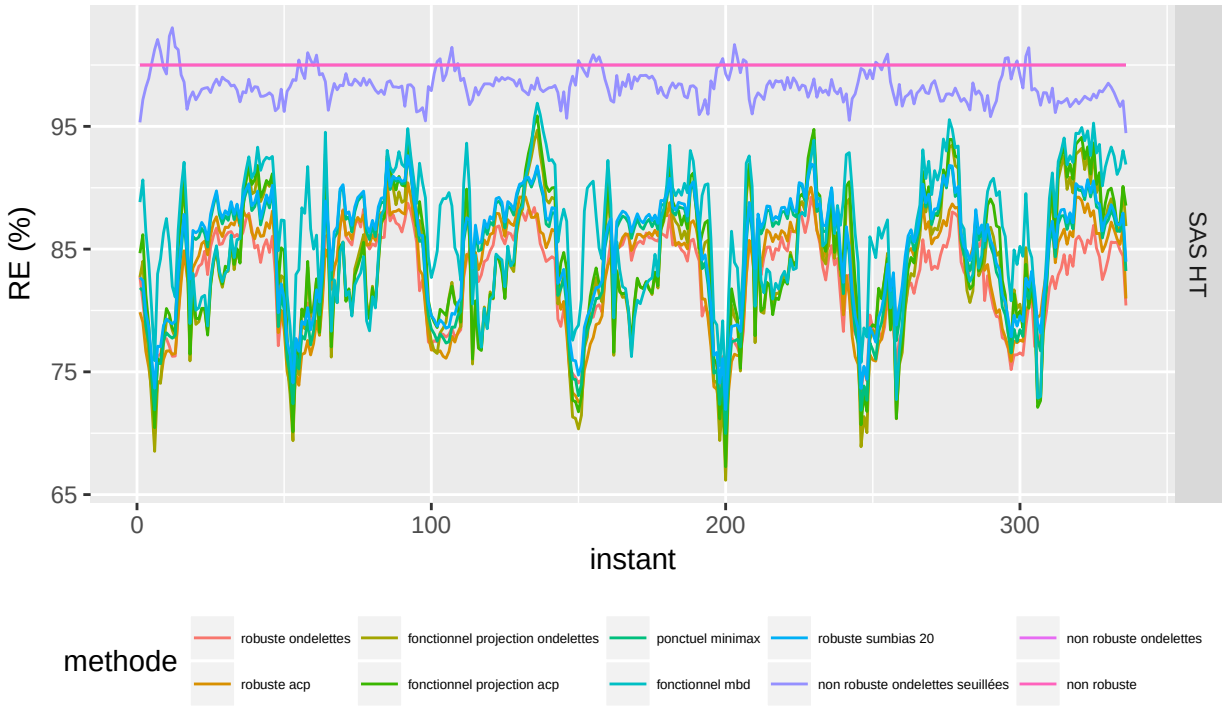

FIGURE 3.15 – Évolution du RE au cours du temps pour les différentes méthodes, Sondage aléatoire simple, et $n = 100$

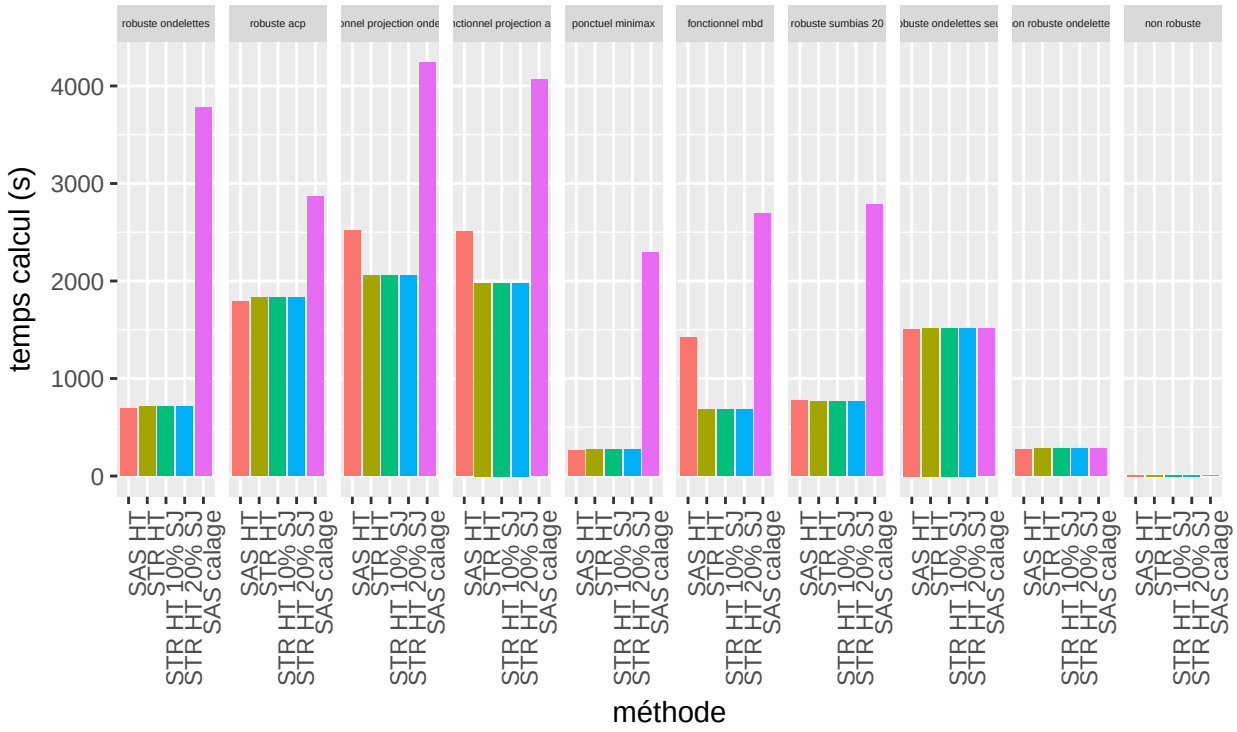

FIGURE 3.16 – Temps de calcul des différents estimateurs (en secondes) en fonction de la taille des échantillons, pour différents plans de sondage.

taille	méthode	RB (%)	RE (%)	tempsCalcul (ms)
40	robuste ondelettes	-7	**72**	105
40	robuste acp	-7	73	80
40	fonctionnel projection ondelettes	-8	73	96
40	fonctionnel projection acp	-8	73	141
40	ponctuel minimax	-9	74	57
40	fonctionnel mbd	-8	75	106
40	robuste sumbias 20	-8	75	99
40	robuste sumbias 10	-7	77	103
40	robuste sumbias 4	-4	85	107
40	non robuste ondelettes seuillées	1	98	83
40	non robuste ondelettes	0	100	17
40	non robuste	0	100	0
100	robuste ondelettes	-4	**83**	127
100	robuste acp	-5	**83**	168
100	fonctionnel projection ondelettes	-5	84	328
100	fonctionnel projection acp	-5	84	288
100	ponctuel minimax	-6	85	63
100	fonctionnel mbd	-5	86	200
100	robuste sumbias 20	-5	85	134
100	robuste sumbias 10	-4	86	138
100	robuste sumbias 4	-2	92	143
100	non robuste ondelettes seuillées	1	98	206
100	non robuste ondelettes	0	100	38
100	non robuste	0	100	0
200	robuste ondelettes	-3	**90**	174
200	robuste acp	-3	**90**	402
200	fonctionnel projection ondelettes	-3	90	688
200	fonctionnel projection acp	-4	91	612
200	ponctuel minimax	-4	92	66
200	fonctionnel mbd	-3	94	401
200	robuste sumbias 20	-4	92	203
200	robuste sumbias 10	-3	92	207
200	robuste sumbias 4	-1	96	200
200	non robuste ondelettes seuillées	0	99	403
200	non robuste ondelettes	-0	100	75
200	non robuste	-0	100	1
400	robuste ondelettes	-2	**94**	292
400	robuste acp	-2	95	1140
400	fonctionnel projection ondelettes	-2	94	1415
400	fonctionnel projection acp	-2	95	1476
400	ponctuel minimax	-3	96	77
400	fonctionnel mbd	-2	97	709
400	robuste sumbias 20	-2	96	341
400	robuste sumbias 10	-2	96	330
400	robuste sumbias 4	-1	98	346
400	non robuste ondelettes seuillées	0	99	817
400	non robuste ondelettes	-0	100	144
400	non robuste	-0	100	4

TABLEAU 3.2 – Indicateurs de qualité pour le sondage aléatoire simple (SAS HT) et l'estimateur de Horvitz-Thompson en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode)

taille	méthode	RB (%)	RE (%)	tempsCalcul (ms)
40	robuste ondelettes	-4	**94**	808
40	robuste acp	-4	98	130
40	fonctionnel projection ondelettes	-5	99	571
40	fonctionnel projection acp	-5	99	541
40	ponctuel minimax	-7	100	526
40	fonctionnel mbd	-5	99	515
40	robuste sumbias 20	-7	100	570
40	robuste sumbias 10	-7	99	565
40	robuste sumbias 4	-6	99	574
40	non robuste ondelettes seuillées	-4	96	85
40	non robuste ondelettes	-4	100	18
40	non robuste	-4	100	2
100	robuste ondelettes	-2	**92**	862
100	robuste acp	-2	94	303
100	fonctionnel projection ondelettes	-2	96	819
100	fonctionnel projection acp	-2	96	731
100	ponctuel minimax	-4	96	550
100	fonctionnel mbd	-2	97	661
100	robuste sumbias 20	-4	96	621
100	robuste sumbias 10	-3	96	618
100	robuste sumbias 4	-2	97	626
100	non robuste ondelettes seuillées	-1	96	204
100	non robuste ondelettes	-1	100	40
100	non robuste	-1	100	2
200	robuste ondelettes	-1	**93**	950
200	robuste acp	-1	94	707
200	fonctionnel projection ondelettes	-1	96	1233
200	fonctionnel projection acp	-1	97	1027
200	ponctuel minimax	-2	97	575
200	fonctionnel mbd	-1	97	701
200	robuste sumbias 20	-2	97	703
200	robuste sumbias 10	-2	97	710
200	robuste sumbias 4	-1	98	720
200	non robuste ondelettes seuillées	-0	97	415
200	non robuste ondelettes	-1	100	76
200	non robuste	-1	100	2
400	robuste ondelettes	-0	**94**	1168
400	robuste acp	-1	95	1732
400	fonctionnel projection ondelettes	-1	98	1615
400	fonctionnel projection acp	-1	98	1765
400	ponctuel minimax	-1	98	645
400	fonctionnel mbd	-1	98	818
400	robuste sumbias 20	-1	98	896
400	robuste sumbias 10	-1	98	888
400	robuste sumbias 4	-1	99	909
400	non robuste ondelettes seuillées	0	98	816
400	non robuste ondelettes	-0	100	151
400	non robuste	-0	100	6

TABLEAU 3.3 – Indicateurs de qualité pour le sondage aléatoire simple avec calage (SAS calage) en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode)

taille	méthode	RB (%)	RE (%)	tempsCalcul (ms)
40	robuste ondelettes	-2	**94**	104
40	robuste acp	-1	97	80
40	fonctionnel projection ondelettes	-0	101	131
40	fonctionnel projection acp	-0	101	103
40	ponctuel minimax	-2	97	61
40	fonctionnel mbd	-0	100	64
40	robuste sumbias 20	-2	97	102
40	robuste sumbias 10	-2	97	104
40	robuste sumbias 4	-1	98	112
40	non robuste ondelettes seuillées	0	95	82
40	non robuste ondelettes	-0	100	17
40	non robuste	-0	100	0
100	robuste ondelettes	-1	**93**	130
100	robuste acp	-1	97	177
100	fonctionnel projection ondelettes	0	100	256
100	fonctionnel projection acp	-0	99	170
100	ponctuel minimax	-2	96	61
100	fonctionnel mbd	-0	99	171
100	robuste sumbias 20	-2	96	133
100	robuste sumbias 10	-2	96	136
100	robuste sumbias 4	-1	97	148
100	non robuste ondelettes seuillées	0	95	204
100	non robuste ondelettes	0	100	41
100	non robuste	0	100	0
200	robuste ondelettes	-0	**94**	178
200	robuste acp	-0	98	411
200	fonctionnel projection ondelettes	0	100	520
200	fonctionnel projection acp	-0	99	420
200	ponctuel minimax	-1	97	67
200	fonctionnel mbd	-0	100	189
200	robuste sumbias 20	-1	97	194
200	robuste sumbias 10	-1	97	196
200	robuste sumbias 4	-0	98	205
200	non robuste ondelettes seuillées	0	97	410
200	non robuste ondelettes	0	100	76
200	non robuste	0	100	1
400	robuste ondelettes	-0	**95**	302
400	robuste acp	-0	98	1159
400	fonctionnel projection ondelettes	-0	101	1158
400	fonctionnel projection acp	-0	99	1290
400	ponctuel minimax	-1	98	91
400	fonctionnel mbd	-0	100	264
400	robuste sumbias 20	-1	98	339
400	robuste sumbias 10	-1	98	337
400	robuste sumbias 4	-0	99	343
400	non robuste ondelettes seuillées	0	100	815
400	non robuste ondelettes	-0	100	153
400	non robuste	-0	100	3

TABLEAU 3.4 – Indicateurs de qualité pour le sondage stratifié sans strata jumper (STR HT) en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode)

taille	méthode	RB (%)	RE (%)	tempsCalcul (ms)
40	robuste ondelettes	-3	**85**	104
40	robuste acp	-3	87	80
40	fonctionnel projection ondelettes	-2	89	131
40	fonctionnel projection acp	-2	88	103
40	ponctuel minimax	-4	89	61
40	fonctionnel mbd	-2	89	64
40	robuste sumbias 20	-4	89	102
40	robuste sumbias 10	-3	90	104
40	robuste sumbias 4	-2	93	112
40	non robuste ondelettes seuillées	0	96	82
40	non robuste ondelettes	-0	100	17
40	non robuste	-0	100	0
100	robuste ondelettes	-2	**83**	130
100	robuste acp	-2	86	177
100	fonctionnel projection ondelettes	-1	86	256
100	fonctionnel projection acp	-1	84	170
100	ponctuel minimax	-3	88	61
100	fonctionnel mbd	-1	87	171
100	robuste sumbias 20	-3	88	133
100	robuste sumbias 10	-2	89	136
100	robuste sumbias 4	-1	94	148
100	non robuste ondelettes seuillées	0	97	204
100	non robuste ondelettes	-0	100	41
100	non robuste	-0	100	0
200	robuste ondelettes	-1	86	178
200	robuste acp	-1	87	411
200	fonctionnel projection ondelettes	-1	86	520
200	fonctionnel projection acp	-1	**85**	420
200	ponctuel minimax	-2	89	67
200	fonctionnel mbd	-1	88	189
200	robuste sumbias 20	-2	90	194
200	robuste sumbias 10	-1	91	196
200	robuste sumbias 4	-1	96	205
200	non robuste ondelettes seuillées	0	98	410
200	non robuste ondelettes	0	100	76
200	non robuste	0	100	1
400	robuste ondelettes	-1	88	302
400	robuste acp	-1	89	1159
400	fonctionnel projection ondelettes	-1	88	1158
400	fonctionnel projection acp	-1	**87**	1290
400	ponctuel minimax	-1	91	91
400	fonctionnel mbd	-1	89	264
400	robuste sumbias 20	-1	91	339
400	robuste sumbias 10	-1	92	337
400	robuste sumbias 4	-0	97	343
400	non robuste ondelettes seuillées	0	99	815
400	non robuste ondelettes	0	100	153
400	non robuste	0	100	3

TABLEAU 3.5 – Indicateurs de qualité pour le sondage stratifié avec 10% de strata jumpers en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode)

taille	méthode	RB (%)	RE (%)	tempsCalcul (ms)
40	robuste ondelettes	-3	**81**	104
40	robuste acp	-3	82	80
40	fonctionnel projection ondelettes	-2	84	131
40	fonctionnel projection acp	-3	83	103
40	ponctuel minimax	-5	84	61
40	fonctionnel mbd	-3	85	64
40	robuste sumbias 20	-4	84	102
40	robuste sumbias 10	-4	85	104
40	robuste sumbias 4	-2	91	112
40	non robuste ondelettes seuillées	0	97	82
40	non robuste ondelettes	-0	100	17
40	non robuste	-0	100	0
100	robuste ondelettes	-3	**82**	130
100	robuste acp	-3	84	177
100	fonctionnel projection ondelettes	-2	83	256
100	fonctionnel projection acp	-2	**82**	170
100	ponctuel minimax	-4	86	61
100	fonctionnel mbd	-2	85	171
100	robuste sumbias 20	-3	87	133
100	robuste sumbias 10	-3	88	136
100	robuste sumbias 4	-1	93	148
100	non robuste ondelettes seuillées	0	97	204
100	non robuste ondelettes	-0	100	41
100	non robuste	-0	100	0
200	robuste ondelettes	-2	86	178
200	robuste acp	-2	87	411
200	fonctionnel projection ondelettes	-1	85	520
200	fonctionnel projection acp	-2	**84**	420
200	ponctuel minimax	-2	88	67
200	fonctionnel mbd	-2	87	189
200	robuste sumbias 20	-2	89	194
200	robuste sumbias 10	-2	90	196
200	robuste sumbias 4	-1	95	205
200	non robuste ondelettes seuillées	1	98	410
200	non robuste ondelettes	0	100	76
200	non robuste	0	100	1
400	robuste ondelettes	-1	89	302
400	robuste acp	-1	90	1159
400	fonctionnel projection ondelettes	-1	88	1158
400	fonctionnel projection acp	-1	88	1290
400	ponctuel minimax	-2	91	91
400	fonctionnel mbd	-1	90	264
400	robuste sumbias 20	-1	92	339
400	robuste sumbias 10	-1	93	337
400	robuste sumbias 4	-0	97	343
400	non robuste ondelettes seuillées	0	99	815
400	non robuste ondelettes	0	100	153
400	non robuste	0	100	3

TABLEAU 3.6 – Indicateurs de qualité pour le sondage stratifié avec 20% de strata jumpers en fonction de la taille d'échantillon (taille) et de l'estimateur (méthode)

méthode Estimation MSE	méthode Estimation Courbe	RB (%)	tempsCalcul (ms)
explicite	non robuste	**-1**	10
généralisé	non robuste	**-1**	6617
populationnel	non robuste	**-1**	12014
explicite	non robuste acp	**-1**	189
explicite	non robuste ondelettes	**-1**	94
explicite	robuste minimax	-28	418
généralisé	robuste minimax	32	16203
populationnel	robuste minimax	3	78037
explicite	robuste minimax acp	-33	331
généralisé	robuste minimax acp	31	8787
explicite	robuste sumbias	-21	513

TABLEAU 3.7 – Biais relatifs et temps de calcul des différentes méthodes d'estimation (sondage aléatoire simple, 100 unités et estimateur de Horvitz-Thompson)

méthode Estimation MSE	méthode Estimation Courbe	RB (%)	tempsCalcul (ms)
explicite	non robuste	**0**	15
généralisé	non robuste	**0**	6755
populationnel	non robuste	-9	18540
explicite	non robuste acp	**0**	169
explicite	non robuste ondelettes	**0**	116
explicite	robuste minimax	-14	417
généralisé	robuste minimax	19	17282
populationnel	robuste minimax	**-6**	84116
explicite	robuste minimax acp	-25	406
généralisé	robuste minimax acp	15	9626
explicite	robuste sumbias	-13	504

TABLEAU 3.8 – Biais relatifs et temps de calcul des différentes méthodes d'estimation (sondage stratifié, 100 unités et estimateur de Horvitz-Thompson)

méthode Estimation MSE	méthode Estimation Courbe	RB (%)	tempsCalcul (ms)
explicite	non robuste	-32	494
généralisé	non robuste	-20	7264
populationnel	non robuste	-21	13663
explicite	non robuste acp	-32	302
explicite	non robuste ondelettes	-32	834
généralisé	robuste minimax	-10	18340
généralisé	robuste minimax acp	-13	9156
généralisé	robuste minimax fonctionnel	223	16907
généralisé	robuste sumbias	-13	16981
explicite	robuste minimax	-47	1327
populationnel	robuste minimax	-28	79154
explicite	robuste minimax acp	-50	604
explicite	robuste sumbias	-46	1375

TABLEAU 3.9 – Biais relatifs et temps de calcul des différentes méthodes d'estimation (sondage aléatoire simple, 100 unités, et estimateur par calage)

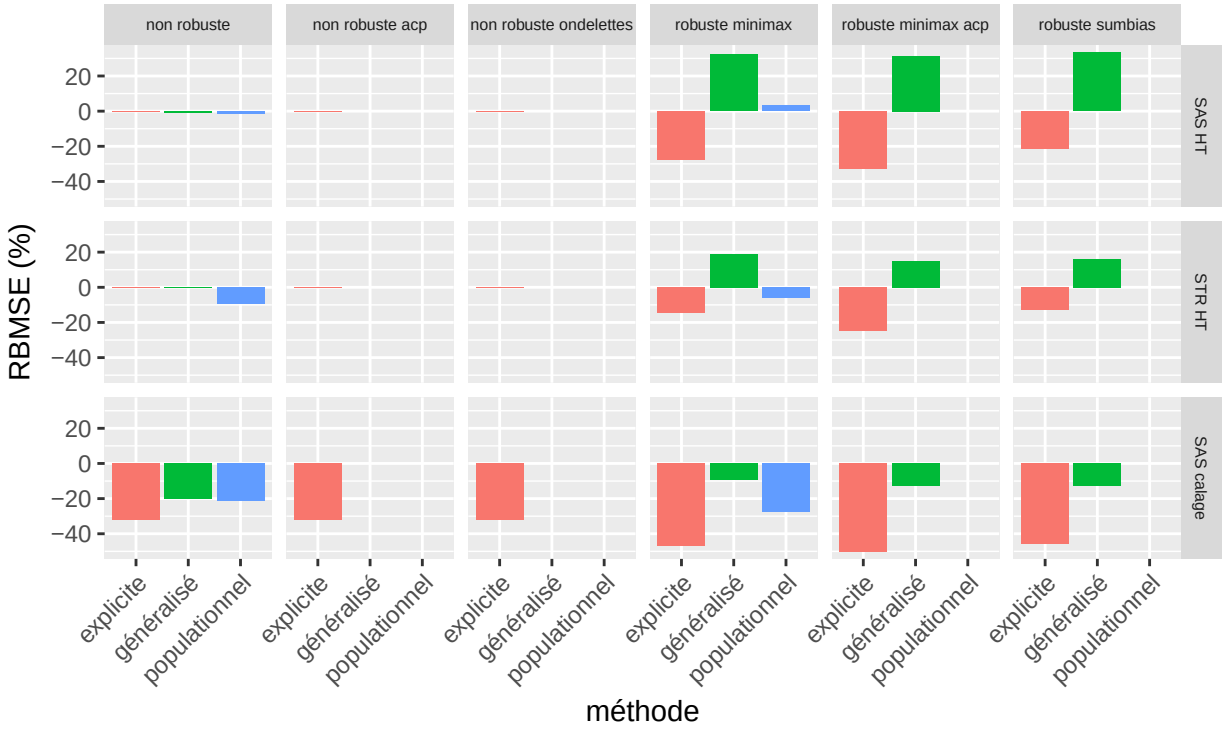

FIGURE 3.17 – Biais relatifs pour les différents estimateurs d'Erreur Quadratique Moyenne et différents estimateurs de courbe moyenne (sondage aléatoire simple, taille d'échantillon $n = 100$)

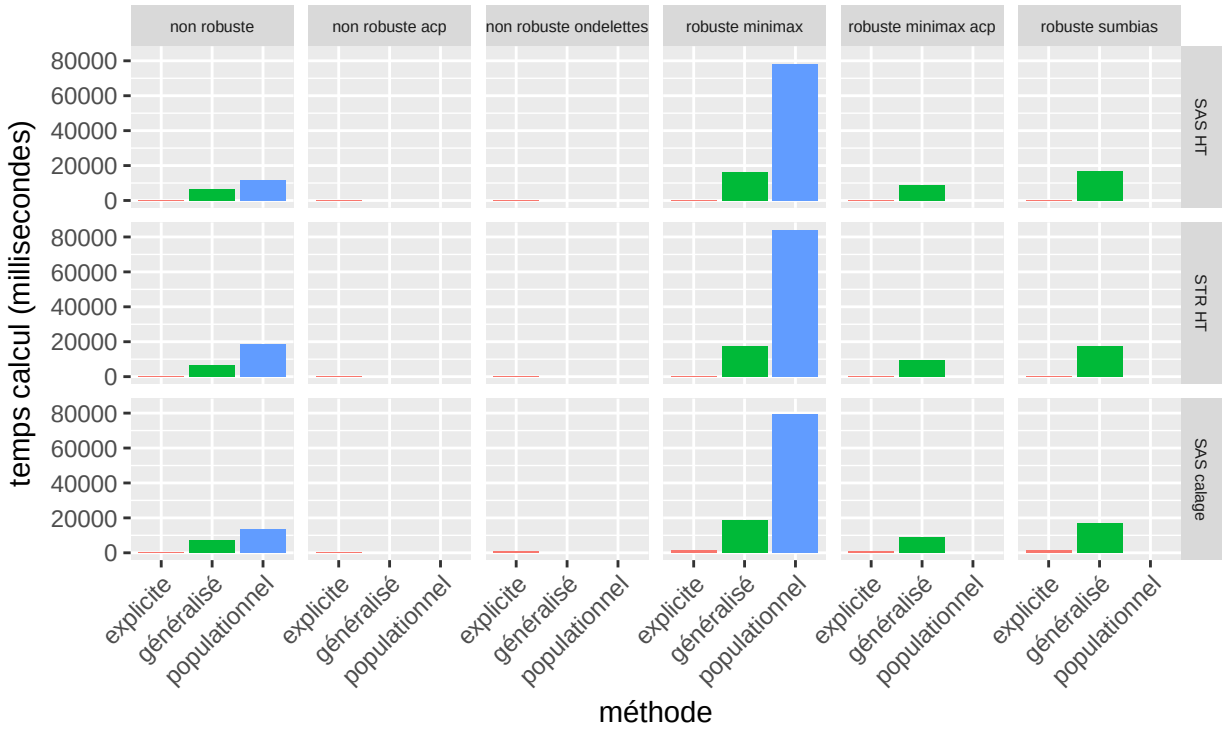

FIGURE 3.18 – Temps de calcul pour les différents estimateurs d'Erreur Quadratique Moyenne et différents estimateurs de courbe moyenne (sondage aléatoire simple, taille d'échantillon $n = 100$)

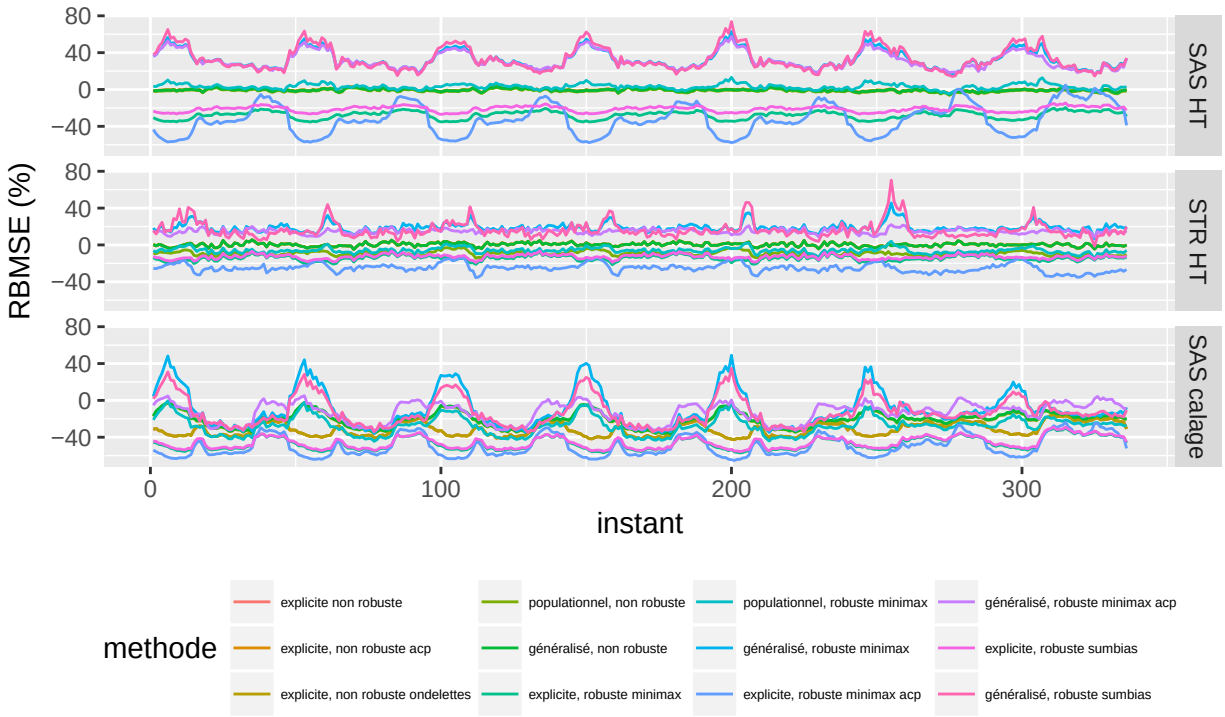

FIGURE 3.19 – Biais relatifs au cours du temps pour les estimateurs d'Erreur Quadratique Moyenne et différents estimateurs de courbe moyenne (sondage aléatoire simple, 100 unités)

3.7 Conclusions sur la robustesse

3.7.1 Conclusions méthodologiques

Dans ce chapitre, nous avons proposé trois approches permettant d'adapter les estimateurs robustes en sondage proposés par Beaumont et al. (2013) au cadre des données fonctionnelles : la première approche consiste simplement à appliquer les méthodes robustes univariées indépendamment sur chacun des instants de discrétisation. Elle ne permet pas de prendre en compte les corrélations temporelles de la problématique. La seconde repose sur le passage d'un problème fonctionnel à un problème de dimension finie par projection sur une base finie de fonctions par exemple une base d'ondelettes ou encore la base des Composantes Principales Sphériques. On estime alors indépendamment les totaux des coordonnées des courbes de la population pour chaque vecteur de cette base selon des méthodes usuelles de construction d'estimateurs robustes pour des variables réelles et enfin on en déduit une estimation de la courbe moyenne. Enfin, la troisième méthode consiste à déterminer de manière fonctionnelle les bornes supérieure et inférieure de la fonction de troncature à appliquer sur les biais conditionnels en dilatant la région centrale contenant les 50 % de courbes les plus "profondes", c'est-à-dire les moins atypiques.

Dans les trois cas, la question centrale est de déterminer la meilleure constante de troncature, qui définit à quel niveau sont bornées les influences individuelles représentées par les biais conditionnels. Pour cela nous proposons une adaptation du critère minimax pour la troncature fonctionnelle, mais aussi un critère alternatif, pouvant se décliner dans les trois approches, conçu pour pouvoir prendre en compte la variabilité de la constante dans les estimations de précision.

Des simulations réalisées sur des données réelles montrent que l'application des méthodes robustes permet des gains de précision notables, en particulier lorsque la variance de l'estimateur non robuste est la plus forte : plans de sondage aléatoires simples, plans de sondages stratifiés en présence de strata jumpers ou encore petits échantillons. Par ailleurs, les méthodes robustes semblent ne jamais dégrader significativement la précision des estimations, même dans les cas où l'estimateur non robuste était de variance modérée. La meilleure des trois méthodes dans les tests est l'estimation par application des méthodes de troncature univariée sur les cœfficients d'ondelettes ou sur les scores de l'Analyse en Composantes Principales Sphériques; elle induit des gains de précision (Erreur Quadratique Moyenne) de 28% pour des échantillons de 40 individus sélectionnés par sondage aléatoire simple, de 17 % pour des échantillons de 100 individus et de 10% pour des échantillons de 200 individus.

En outre, nous définissons également différentes méthodes d'estimation d'erreur quadratique moyenne pour nos différents estimateurs. Une première approche consiste à considérer les constantes de troncature comme fixes bien qu'elles dépendent en réalité de l'échantillon. Pour des plans de sondage stratifiés, on peut alors proposer des estimations explicites d'erreur quadratique moyenne, rapides à mettre en œuvre. Néanmoins, cela induit une sous-estimation notable de l'erreur de l'ordre de 30 %.

Pour limiter cette sous-estimation et prendre en compte la variabilité de la constante, on propose donc d'utiliser des bootstraps, populationnels ou généralisés,

en considérant des constantes de troncature choisies à partir des données pour chaque rééchantillon. Le bootstrap populationnel semble donner des résultats satisfaisants, avec un biais relatif dont la valeur absolue est entre 3% et 6% pour les estimateurs de Horvitz-Thompson. Les bootstraps sont cependant sensiblement plus longs à mettre en œuvre que les estimations explicites (80 secondes contre moins d'une seconde pour un échantillon de 100 courbes et 336 points de mesure).

3.7.2 Cas d'application des méthodes robustes

Les méthodes proposées peuvent engendrer des gains de précision notables lorsque l'on s'intéresse à une ou plusieurs sous-populations de tailles réduites. Parmi les trois approches étudiées, on préconise de privilégier l'approche par projection sur la base de l'ACP robuste ou encore la projection sur les bases d'ondelettes, qui semblent dans les tests offrir les meilleures performances et qui de plus permet de préserver les corrélations temporelles de la problématique.

On utilisera ces méthodes lorsque l'on s'intéresse à une sous-population unique, ou à plusieurs sous-populations pour lesquelles il n'est pas raisonnable de postuler un modèle de superpopulation commun (un exemple étant celui des secteurs d'activités ou des modes de chauffage et un contre-exemple étant celui des zones géographiques). Dans le cas contraire, on préférera utiliser les méthodes d'estimation sur petits domaines présentés dans le chapitre suivant afin de tirer parti des similitudes entre les courbes des différents domaines pour rendre plus robustes les estimations sur chacun.

3.7.3 Perspectives

De nombreuses questions restent encore ouvertes sur ce sujet de l'estimation robuste pour des courbes moyennes ou totales. Nous n'avons notamment pas étudié la question de la fenêtre temporelle à prendre en compte. En effet, il est fréquent que l'on ait à travailler sur de très longues périodes temporelles (plusieurs années au pas demi-horaire). Si l'on souhaite utiliser une méthode robuste basée sur l'ACP sur les instants de discrétisation, il n'est alors pas pertinent de travailler sur l'ensemble de la courbe (il y aurait alors beaucoup plus d'instants que d'individus et beaucoup de valeurs propres nulles dans l'analyse spectrale). Il serait alors judicieux de découper notre période en plus petites périodes qui seront traitées séparément (semaines ou mois). On peut alors se demander comment choisir la fenêtre temporelle optimale et comment traiter les estimations en bordures de ces fenêtres. De même, pour les méthodes fonctionnelles, la taille de la fenêtre temporelle sélectionnée aura un impact important sur les limites de troncature obtenues.

Par ailleurs, nous n'avons proposé ici que des estimateurs instantanés de l'erreur quadratique moyenne. Or il pourrait être intéressant de construire des bandes de confiance qui prennent en compte les corrélations temporelles du problème : une bande de confiance à α % d'un estimateur de courbe moyenne ou totale se définit comme l'enveloppe dans laquelle la courbe moyenne ou totale de la population est intégralement contenue avec une probabilité de α %, pour $\alpha \in [0,1]$. Pour construire ces bandes de confiance, on pourrait se baser sur la démarche proposée par Cardot et al. (2013b) dans le cadre de l'estimation non robuste et l'adapter à notre contexte de

l'estimation robuste dans laquelle nous avons non seulement de la variance mais aussi du biais.

Chapitre 4

Estimation de courbes de consommation électrique moyennes ou totales par sondage pour de petits domaines

Dans ce chapitre, nous nous intéressons à l'estimation de courbes de consommation électrique moyennes ou totales pour de petites sous-populations, aussi appelées petits domaines.

4.1 Contexte et introduction

Avec l'arrivée des compteurs communicants, il deviendra de plus en plus aisé et de moins en moins coûteux de recruter et maintenir de grands échantillons de courbes de charge. Il sera donc possible de produire des estimations de courbes moyennes non plus uniquement à la maille de l'ensemble de la France, mais aussi pour de plus petites zones géographiques telles que les régions, les départements, les villes, voire même les quartiers. Ces estimations pourraient être utilisées par exemple pour proposer aux collectivités territoriales des services basés sur l'analyse de leur courbe de consommation électrique estimée, ou encore pour aider Enedis à assurer l'équilibre entre offre et demande d'énergie au niveau local et en particulier à intégrer les énergies renouvelables, actuellement en plein développement, sur le réseau. En outre, dans un souci de développement de l'Open Data, Enedis pourrait souhaiter publier des informations sur les consommations électriques à des mailles les plus fines possibles géographiquement et temporellement, et dans ce cadre, une solution pourrait être d'exploiter les résultats de nos estimations.

Par ailleurs, les modèles développés dans ce chapitre, qui décrivent le lien entre la courbe de consommation électrique individuelle et les variables explicatives, peuvent être appliqués dans d'autres contextes que l'estimation de courbes moyennes de sous-populations. On peut en effet en faire usage dans une optique de prospective, pour prévoir l'évolution de la consommation globale d'un territoire en fonction des prévisions d'évolution de différents paramètres : par exemple, si nous avons identifié dans nos panels des clients équipés de véhicules électriques, les modèles peuvent nous per-

mettre de quantifier l'impact de ce nouvel usage sur la courbe de charge et donc d'anticiper l'effet de son développement sur la courbe de charge globale. Enfin, les modèles peuvent également permettre de proposer un service au client final basé sur la comparaison de sa courbe de consommation à celle d'un client-type possédant des caractéristiques similaires, estimée par le modèle.

Bien que le besoin d'estimation pour de petites sous-populations soit grandissant, nos panels ont été dimensionnés pour assurer des précisions correctes à l'échelle de la France entière, et non de ces domaines. Si les sous-populations d'intérêt sont petites, on n'aura donc que peu d'unités dans chacune d'entre elles, voire pas du tout dans certains cas. Cela rendra les estimations classiques directes -c'est-à-dire basées uniquement sur les données du domaine considéré - imprécises, voire impossibles à réaliser si aucune unité du domaine n'est échantillonnée.

En sondages, cette question est très fréquemment abordée, et connue sous le terme d'estimation sur de petits domaines. Un "petit" domaine peut ici se définir comme un domaine pour lequel on ne dispose que de peu d'unités dans l'échantillon *au regard de ce qu'on souhaite mesurer*, c'est-à-dire un domaine pour lequel l'estimation est imprécise. Pour améliorer la qualité des estimations produites sur les petits domaines, les méthodes de l'état de l'art reposent souvent sur des modélisations implicites ou explicites utilisant conjointement les données de l'ensemble des domaines. En accroissant la quantité d'information exploitée, les modèles permettent de consolider les estimations produites sur chacun des domaines. Le livre récent de Rao and Molina (2015) propose un état de l'art des méthodes existantes.

Lorsque la variable d'intérêt est réelle, les méthodes les plus couramment utilisées pour l'estimation sur des petits domaines sont le modèle linéaire mixte au niveau domaine proposé par Fay III and Herriot (1979) et le modèle linéaire mixte au niveau unité proposé par Battese et al. (1988). Dans notre contexte, on dispose principalement de données au niveau unité, c'est donc ce dernier modèle qu'on se proposera d'adapter au contexte fonctionnel dans la sous section 4.3.1.

Dans la littérature, il existe des méthodes d'estimation sur petits domaines spécifiques aux séries temporelles. Par exemple, Burck and Pfeffermann (1990) et Rao and Yu (1994) posent des modèles de séries temporelles sur les séries d'aléas et/ou de cœfficients aux différents instants pour prendre en compte les dépendances temporelles. Cependant ces modèles de type espace-état ont été pensés pour des séries temporelles relativement courtes (quelques dizaines de points). Ils s'estiment en effet par des filtres de Kalman, très gourmands en temps de calcul. De plus, nos courbes sont très chahutées, et présentent une structure temporelle très complexe avec de nombreuses saisonnalités ; certains outils d'analyse des données fonctionnelles tels que les ondelettes apparaissent donc plus pertinents que les méthodes de séries temporelles de type ARIMA.

Toujours pour des raisons de temps de calculs et de rapidité d'implémentation, nous avons privilégié les modèles linéaires mais on aurait également pu utiliser d'autres approches. Ainsi, Opsomer et al. (2008) propose une extension non linéaire des modèles mixtes au niveau unité à l'aide de P-splines. Cette approche est très intéressante mais néanmoins coûteuse en temps de calcul, ce qui peut être problématique dans un cas d'application tel que le nôtre dans lequel les données peuvent être volu-

mineuses. [1]

Nous proposons des adaptations de méthodes existantes pour l'estimation sur petits domaines au contexte des données fonctionnelles. Dans un premier temps, dans la section 4.2 on se place dans l'approche basée sur le plan de sondage, et les estimateurs utilisés sont l'estimateur de Horvitz-Thompson et l'estimateur par calage de courbe moyenne présentés respectivement dans (2.14) et (2.23), appliqué indépendamment sur chaque domaine. Comme la courbe moyenne de chaque domaine est estimée séparément, cette approche n'exploite pas conjointement l'ensemble des données, c'est pourquoi dans la section 4.3, on supposera l'existence d'un modèle commun à l'ensemble des unités de la population, liant l'information auxiliaire et la variable d'intérêt, et qui permettent, en utilisant conjointement les données issues de l'ensemble des domaines, de gagner en précision. On se placera alors dans le cadre de l'estimation basée sur un modèle.

Notre première approche, exposée dans la sous-section 4.3.1, s'inscrit dans une philosophie similaire à celle suivie dans le précédent Chapitre. Elle consiste à projeter notre problème fonctionnel dans un espace de dimension finie puis à appliquer indépendamment sur chacun des vecteurs de base de cet espace une des méthodes les plus usuelles de la littérature sur la thématique de l'estimation pour de petits domaines pour des variables scalaires. En l'occurrence, il s'agira ici des modèles linéaires mixtes au niveau unité proposés par Battese et al. (1988). Comme dans le Chapitre précédent, on testera différents espace de projection, en particulier les composantes principales de l'ACP et les bases d'ondelettes.

Une seconde approche, exposée dans la sous-section 4.3.2, consiste à utiliser des régressions linéaires fonctionnelles. Dans le cadre de l'estimation basée sur le plan de sondage, et lorsque l'on peut faire l'hypothèse que la distribution de la variable étudiée est la même dans l'ensemble des domaines conditionnellement aux informations auxiliaires, Ardilly (2014) propose d'estimer les courbes de chacun des domaines aisément grâce au GREG (Generalized Regression Estimator) étudié par Särndal (1992), qui est un estimateur assisté par un modèle. Cette méthode possède l'immense avantage de ne nécessiter que l'estimation d'un seul poids par unité même lorsque l'on s'intéresse à un grand nombre de variables, en l'occurrence les valeurs de la courbe à chacun des instants de la période considérée. Nous montrons donc dans la sous-section 4.3.2 comment adapter cette démarche dans notre cadre de travail où l'estimation est basée sur un modèle et non pas sur le plan de sondage.

Une troisième famille de méthodes, présentée dans la sous-section 4.3.3, consiste à suivre l'approche par prédiction décrite dans Valliant et al. (2000) : il s'agit de construire un estimateur de la courbe moyenne en agrégeant les prédictions individuelles de chacune des unités non échantillonnées et les courbes échantillonnées. Pour produire ces prédictions individuelles, nous utilisons des arbres de régression adaptées au contexte des données fonctionnelles selon l'approche dite du "Courbotree", fréquemment utilisée à EDF et décrite dans Stéphan and Cogordan (2009). Nous

1. Le concept de ce chapitre, qui est d'associer dans une même problématique petits domaines et données massives, peut de prime abord sembler contre-intuitif mais c'est néanmoins la situation à laquelle nous sommes confrontés : on s'intéresse en effet à des *petits domaines* nombreux, pour lesquels on souhaite estimer la courbe moyenne à de nombreux instants, et, pour chacun de ces domaines, on ne dispose que de très peu voire pas du tout de données.

avons étendu cette approche de façon à utiliser des forêts aléatoires plutôt que des arbres de régression. Nous aurions cependant tout à fait pu utiliser d'autres algorithmes d'analyse des données fonctionnelles.

Comme dans le Chapitre précédent, nos estimations de courbes moyennes ou totales de petits domaines peuvent être sensibles à la présence dans les échantillons d'unités influentes. On se propose donc dans la section 4.4 d'aborder conjointement les aspects petites sous-populations et unités influentes en construisant des estimateurs robustes de courbes moyennes ou totales pour de petits domaines. Cette question de l'estimation robuste en sondages pour des petits domaines a déjà été abordée dans la littérature, hors du contexte des données fonctionnelles. Ainsi, pour des variables d'intérêt réelles, Jiongo et al. (2013) propose une approche basée sur le biais conditionnel, similaire à la nôtre, dans le cadre de modèles linéaires mixtes au niveau unité, que nous appliquons ici pour estimer les coordonnées moyennes de chaque domaine pour chaque vecteur de la base de projection dans le paragraphe 4.4.1.

Dans la littérature, il existe également des méthodes d'estimation robuste sur des petits domaines qui ne se basent pas sur la notion de biais conditionnels. Ainsi, Sinha and Rao (2009) proposent l'estimateur REBLUP (Robust EBLUP) qui est une version robuste du modèle linéaire mixte niveau unité obtenue en introduisant des fonctions de Huber dans les équations d'estimation du maximum de vraisemblance utilisées pour estimer les paramètres. L'estimateur obtenu appartient alors à la famille des M-estimateurs. L'une des méthodes que nous proposons, détaillée au paragraphe 4.4.2, consiste donc à déployer cette démarche sur les vecteurs d'une base de projection des courbes. Une version semi-paramétrique du REBLUP a également été proposée par Rao et al. (2014) à partir d'un modèle de P-splines.

Par ailleurs, Tzavidis et al. (2010) et Chambers and Tzavidis (2006) utilisent la méthode des M-quantiles pour répondre à cette même problématique. Leur proposition se base sur des régressions quantiles robustes (M-quantiles) appliquées pour estimer la courbe moyenne de chaque domaine en fonction du vecteur des moyennes des variables explicatives, pour un certain quantile α déterminé à partir des unités échantillonnées du domaine et qui permet de capter les spécificités du domaine non traduites dans les variables explicatives. Cette méthode est cependant assez gourmande en temps de calcul, c'est pourquoi nous ne l'avons pas retenue dans nos tests.

L'ensemble des méthodes proposées dans ce chapitre sont finalement testées et comparées sur un jeu de données de courbes de consommation électrique de ménages français dans la section 4.5.

4.1.1 Notations et cadre de travail

On se place dans le cadre de travail défini dans le paragraphe 2.2.1. Dans cette sous-section, nous introduisons quelques notations additionnelles, notamment sur la définition des domaines et sur les variables explicatives.

La population U peut être décomposée en D domaines disjoints $U_1, \dots, U_d, \dots, U_D$ de tailles $N_1, \dots, N_D$ connues. Notre but est d'estimer la courbe moyenne μ_d de chaque domaine, *i.e.*

$$\mu_d(t) = \frac{1}{N_d} \sum_{i \in U_d} Y_i(t), t \in [0, T]. \tag{4.1}$$

Dans la population U on sélectionne un échantillon s de taille n selon un plan de sondage aléatoire. On note $s_d = \mathrm{U}_d \cap s$ l'intersection du domaine U_d et de l'échantillon s et n_d la taille de s_d. La taille n_d, supposée connue, est aléatoire et peut être égale à 0 pour un ou plusieurs domaines.

A EDF, on dispose en général d'informations auxiliaires pertinentes, tant au niveau individuel qu'à la maille des domaines étudiés. Ainsi, au niveau individuel, pour chacun des clients de la population, on dispose de données de facturation, qui nous renseignent sur l'énergie totale consommée par le client en un an, mais aussi sur sa puissance souscrite, son tarif ainsi que ses plages d'heures creuses éventuelles. On sait parfois également si le client vit en appartement ou en maison individuelle. Dans un futur proche, on connaîtra le code NAF de chacun des clients professionnels, qui décrit son secteur d'activité. On déduit également du code postal du client la zone climatique dans laquelle il vit et le caractère rural ou urbain de la commune de résidence. Enfin, on connaît le code IRIS de chacun des clients. Les IRIS sont un découpage géographique très fin de la France, chacune de ces zones contenant environ 1500 personnes. Pour chacune de ces IRIS, l'INSEE propose en Open Data sur son site `http://www.insee.fr/` des informations socio-démographiques et économiques complètes. Dans notre contexte d'étude des courbes de consommation électrique, on pourra se servir en particulier de la répartition des résidences principales en fonction de leur taille, des taux de résidences secondaires et des taux d'équipements en chauffage électrique. On va donc utiliser ces données dans nos modèles.

On dispose donc d'un vecteur d'informations auxiliaires $\mathbf{X}_i^0$ connu pour chaque individu i de la population ainsi que d'informations auxiliaires complémentaires, cette fois au niveau des domaines, Z_d. Pour appliquer directement les méthodes linéaires, on regroupe ces informations dans le nouveau vecteur $\mathbf{X}_i = (\mathbf{X}_i^0, \mathrm{Z}_s) \forall i \in \mathrm{U}_d$. On connaît également les moyennes $\overline{\mathbf{X}}_d$ des variables $\mathbf{X}_i$ pour chaque domaine. Pour simplifier, on considère ici que ces variables explicatives sont sans valeur manquante et constantes au cours du temps.

4.2 Estimation de courbes moyennes pour des petits domaines, dans l'approche basée sur le plan de sondage

Dans cette section, on se place dans l'approche basée sur le plan de sondage. Cela signifie que l'on considère que les valeurs de la variable d'intérêt Y_i pour chaque unité de la population sont déterministes et que le seul aléa présent est celui de la constitution de l'échantillon. L'inférence statistique décrit alors uniquement le hasard engendré par le plan de sondage.

Nous allons utiliser deux estimateurs classiques, l'estimateur de Horvitz-Thompson et l'estimateur par calage. Il s'agit d'estimateurs directs, c'est-à-dire des estimateurs construits en n'utilisant, pour l'estimation de la moyenne de chaque domaine, que les unités et les informations auxiliaires relatives au domaine concerné.

Le premier estimateur est l'estimateur de Horvitz-Thompson (Horvitz and Thompson (1952)), qui, étendu au cadre fonctionnel (voir Cardot et al. (2013b)) et pour un

domaine d, s'écrit

$$\hat{\mu}_d^{\mathrm{HT}}(t) = \frac{1}{\mathrm{N}_d} \sum_{i \in s_d} d_i \mathrm{Y}_i(t), \quad d = 1, \ldots \mathrm{D}, \quad t \in [0, \mathrm{T}], \tag{4.2}$$

avec $d_i = \frac{1}{\pi_i}$ le poids de sondage de l'unité i, aussi appelé poids de Horvitz-Thompson.

Cet estimateur très simple nous servira de référence afin d'évaluer les performances de nos méthodes. Il ne peut évidemment pas être calculé pour les domaines non échantillonnés (i.e. les domaines d tels que s_d est vide) et il est extrêmement instable pour les domaines de petite taille. En outre, il ne tire aucunement parti des variables explicatives à notre disposition.

Pour exploiter les informations auxiliaires, toujours dans l'approche basée sur le plan de sondage, on peut utiliser l'estimateur par calage déjà introduit dans la section 2.2.4. Plus formellement, dans le cadre des données fonctionnelles et pour un domaine d, cet estimateur est donné par

$$\hat{\mu}_d^{\mathrm{CALDIR}} = \frac{\sum_{i \in s_d} w_{i s_d} \mathrm{Y}_i}{\mathrm{N}_d}, \tag{4.3}$$

avec les poids $w_{i s_d}$ les plus proches possibles des poids de sondage $d_i = \frac{1}{\pi_i}$ des unités de s_d qui respectent la contrainte

$$t_{\mathbf{X}_d} = \sum_{i \in \mathrm{U}_d} \mathbf{X}_i = \sum_{i \in s_d} w_{i s_d} \mathbf{X}_i.$$

Ces poids $w_{i s_d}$ sont donc les solutions du programme de minimisation :

$$\min_{w_i} \sum_{i \in s_d} d_i \mathrm{G}(w_i, d_i) \quad \text{sous la contrainte} \quad \sum_{i \in s_d} w_i \mathbf{X}_i = \sum_{i \in \mathrm{U}_d} \mathbf{X}_i = t_{\mathbf{X}_d}. \tag{4.4}$$

La résolution de ce problème se fait à l'aide de multiplicateurs de Lagrange.

Cet estimateur, pour chacun des domaines, ne se base que sur les données du domaine concerné (courbes et variables explicatives) sans tenir compte du reste de l'échantillon. Tout comme l'estimateur de Horvitz-Thompson, il est donc imprécis pour les petits domaines et ne peut pas être calculé pour les domaines non échantillonnés.

Les méthodes que nous présentons dans la section suivante vont nous permettre, en posant un modèle commun à l'ensemble des unités de la population qui décrit le lien entre variable d'intérêt et informations auxiliaires, d'exploiter conjointement l'ensemble des données de l'échantillon pour réaliser l'estimation de chacun des domaines, et donc de gagner en précision sur chacun d'entre eux. En outre, cela permettra de pouvoir fournir des estimations même pour les domaines non échantillonnées.

4.3 Estimation de courbes moyennes pour des petits domaines dans l'approche basée sur un modèle

Dans cette section, nous présentons quatre approches innovantes permettant de répondre à notre problématique d'estimation de courbes de charge moyennes sur de

petits domaines. Nous nous plaçons maintenant dans l'approche basée sur le modèle, décrite notamment dans Valliant et al. (2000). Dans cette approche contrairement à l'approche basée sur le plan de sondage dans laquelle s'inscrivait la section précédente, les courbes Y_i ne sont pas considérées comme fixes mais aléatoires. La qualité de nos estimations dépend alors de la qualité de notre modèle : si le modèle est faux, cela pourra conduire à des biais dans les estimations.

On suppose que les informations auxiliaires sont liées aux courbes de charge selon un modèle, dit de superpopulation, valide sur l'ensemble de la population, et qui s'écrit de manière générale

$$Y_i(t) = f_d(\mathbf{X}_i, t) + \epsilon_i(t), \quad i \in U_d, \quad d = 1, \ldots, D, \quad t \in [0, T], \tag{4.5}$$

avec f_d une fonction de régression inconnue à estimer, qui peut varier d'un domaine à l'autre et ϵ_i un processus de bruit d'espérance nulle, de covariance nulle pour des individus différents et non-nulle par rapport au temps.

Dans chacune des quatre sous-sections suivantes, nous estimerons une forme particulière du modèle général (4.5) : les modèles linéaires mixtes au niveau unité appliqués aux scores de l'Analyse en Composantes Principales dans la sous-section 4.3.1, l'estimateur par la régression linéaire fonctionnelle dans la sous-section 4.3.2 puis les arbres de régression pour données fonctionnelles dans la sous-section 4.3.3 et enfin les forêts aléatoires pour données fonctionnelles dans la sous-section 4.3.4.

Comme nous le verrons, chacune de ces méthodes a ses avantages et ses inconvénients comparatifs et le choix de la méthode la plus pertinente dépendra donc du problème traité.

4.3.1 Modèles linéaires mixtes au niveau unité pour des données fonctionnelles

Les modèles linéaires mixtes au niveau unité proposés par Battese et al. (1988) sont très utilisés dans le cadre de l'estimation de totaux de variables réelles pour des domaines. En effet, il permettent, comme nous le verrons plus en détail par la suite, de traduire à la fois l'effet de l'information auxiliaire sur la variable d'intérêt (par les effets fixes) et les spécificités des domaines (par les effets aléatoires).

Dans ce chapitre, nous cherchons donc à adapter ces modèles au contexte des données fonctionnelles. Pour cela, nous allons transformer notre problème fonctionnel en plusieurs sous-problèmes d'estimation de totaux ou de moyennes de variables réelles non corrélées sur des petits domaines, que nous résoudrons ensuite par des méthodes usuelles. Nous proposons donc de mettre en œuvre une approche en trois temps. La première étape consiste à projeter les courbes de notre échantillon dans un espace de dimension finie. Dans la seconde étape, nous estimons pour chaque domaine de la population les moyennes des coordonnées des courbes dans la base de projection. Pour cela, nous appliquons des modèles linéaires mixtes au niveau unité pour des variables réelles. Enfin dans la troisième étape nous combinons les résultats obtenus précédemment en reconstituant les courbes moyennes estimées à partir des moyennes des coordonnées dans l'espace de projection.

L'utilisation de bases de projection permet donc de préserver la structure de corrélation temporelle de nos données tout en se ramenant à plusieurs sous-problèmes décorrélés d'estimation de moyennes de variables réelles que l'on traite indépendamment par les méthodes usuelles.

Plus précisément, nous utilisons ici une ACP fonctionnelle déjà présentée dans 2.1.3. En suivant l'expansion de Karhunen-Loève, chaque courbe de la population peut donc être approximée selon l'équation (2.4).

Remarque 8. *Lorsque nous traiterons la question de l'estimation robuste pour des petits domaines, nous utiliserons plutôt la médiane spatiale m et les composantes principales sphériques de l'ACP robuste* $Z_k, \quad k = 1, \ldots K$, *introduites dans la sous-section 3.4.2 plutôt que la moyenne de l'espace* μ *et les composantes principales* $\zeta_k, \quad k = 1, \ldots K$.

Voyons maintenant comment utiliser l'ACP fonctionnelle pour construire nos estimateurs de courbes moyennes en adaptant les méthodes standards d'estimation de totaux de variables réelles pour des petits domaines. En utilisant (2.4), la moyenne μ_d sur le domaine d peut être approximée par

$$\mu_d(t) \simeq \mu_Y(t) + \sum_{k=1}^{K} \left(\frac{1}{N_d} \sum_{i \in U_d} g_{ik} \right) \zeta_k(t), \quad d = 1, \ldots, D, \quad t \in [0, T]. \tag{4.6}$$

La moyenne inconnue μ_Y est estimée par

$$\hat{\mu}(t) = \frac{1}{N} \sum_{i \in s} d_i Y_i(t), \quad t \in [0, T] \tag{4.7}$$

et les $\zeta_k, k = 1, \ldots, K$ sont estimés par $\hat{\zeta}_k$, les vecteurs propres de $\hat{V}$, l'estimateur de V (Cardot et al. (2010)) :

$$\hat{V}\hat{\zeta}_k = \hat{\lambda}_k \hat{\zeta}_k, \tag{4.8}$$

avec $\hat{\lambda}_k$ l'estimateur de λ_k. Donc, afin d'estimer μ_d, il nous faut estimer la moyenne des scores sur les composantes principales pour le domaine d, *i.e.* $\overline{g}_{dk} = \frac{1}{N_d} \sum_{i \in U_d} g_{ik}$.

Pour cela, nous considérons pour chaque composante $k = 1, \ldots, K$ un modèle linéaire mixte au niveau unité sur g_{ik} (Rao and Molina (2015), Chapitre 4) :

$$g_{ik} = \mathbf{B}_k^{'} \mathbf{X}_i + \upsilon_{dk} + \varepsilon_{ik}, \quad i \in U_d, \quad k = 1, \ldots, K, \tag{4.9}$$

avec $\mathbf{B}_k^{'} \mathbf{X}_i$ l'effet fixe des informations auxiliaires, υ_{dk} l'effet aléatoire du domaine d et ε_{ik} le résidu de l'unité i. On suppose que les effets aléatoires des domaines sont indépendants, et suivent une loi commune de moyenne 0 et de variance $\sigma^2_{\upsilon k}$. Les résidus sont également indépendants, distribués selon une loi de moyenne 0 et de variance $\sigma^2_{\varepsilon k}$. En outre, les effets aléatoires et les résidus sont également supposés indépendants.

Ce modèle, aussi appelé *nested error regression model*, a été introduit par Battese et al. (1988). C'est un modèle paramétrique dans lequel on impose que les effets des variables auxiliaires soient linéaires. Il s'agit d'un modèle linéaire mixte standard, dont les paramètres peuvent être estimés par un BLUP (Best Linear Unbiased Estimator), en suivant l'approche de Rao and Molina (2015), Chapitre 7.

Plus précisément, soient $\overline{\mathbf{X}}_{d,s} = \frac{1}{n_d}\sum_{i\in s_d}\mathbf{X}_i$ et $\overline{g}_{dk,s} = \frac{1}{n_d}\sum_{i\in s_d} g_{ik}$ les moyennes respectives des vecteurs $\mathbf{X}_i$ et des scores g_{ik} sur s_d. Soit $\tilde{\mathbf{B}}_k$ l'estimateur BLUP de $\mathbf{B}_k$.

L'estimateur BLUP de $\overline{g}_{dk}$ s'écrit comme un estimateur composite (voir Rao and Molina (2015)) :

$$\tilde{\overline{g}}_{dk} = \gamma_k\left(\overline{g}_{dk,s} - (\overline{\mathbf{X}}_{d,s} - \overline{\mathbf{X}}_d)'\tilde{\mathbf{B}}_k\right) + (1-\gamma_k)\overline{\mathbf{X}}_d'\tilde{\mathbf{B}}_k, \quad k = 1,\ldots,\mathrm{K}, \tag{4.10}$$

avec

$$\gamma_k = \frac{\sigma^2_{\upsilon k}}{\sigma^2_{\upsilon k} + \sigma^2_{\varepsilon k}}, \quad k = 1,\ldots\mathrm{K}. \tag{4.11}$$

Le premier terme de (4.10) dépend principalement de termes calculés sur s_d : la moyenne des scores et des effets fixes des informations auxiliaires pour s_d. Il reflète les spécificités des scores g_{ik} des unités du domaine d non explicables par les informations auxiliaires. L'estimateur $\hat{\overline{g}}_{dk}$ est donc une combinaison linéaire d'un estimateur traduisant les particularités du domaine d par rapport aux autres et de l'estimateur basé sur un modèle $\overline{\mathbf{X}}_d'\tilde{\mathrm{B}}_k$. L'importance relative de ces deux termes dépendra des variances respectives des effets aléatoires et des résidus. En effet, plus la variance des effets aléatoires grandit, plus il existe de différences entre les domaines non expliquées par l'information auxiliaire, et plus on attribuera un poids fort au premier terme de l'expression, qui justement représente ces différences.

Les variances $\sigma^2_{\upsilon k}$ et $\sigma^2_{\varepsilon k}$ pour $k = 1,\ldots\mathrm{K}$ sont inconnues et elles sont estimées par $\hat{\sigma}^2_{\upsilon k}$ et $\hat{\sigma}^2_{\varepsilon k}$ obtenues par exemple par maximum de vraisemblance restreint (Rao and Molina (2015)). L'estimateur du $\overline{g}_{dk}$ obtenu ainsi est appelé EBLUP (Empirical Best Linear Unbiased Prediction) et donné par

$$\hat{\overline{g}}_{dk} = \hat{\gamma}_k\left(\overline{g}_{dk,s} - (\overline{\mathbf{X}}_{d,s} - \overline{\mathbf{X}}_d)'\tilde{\mathbf{B}}_k\right) + (1-\hat{\gamma}_k)\overline{\mathbf{X}}_d'\tilde{\mathbf{B}}_k, \quad k = 1,\ldots\mathrm{K}. \tag{4.12}$$

où

$$\hat{\gamma}_k = \frac{\hat{\sigma}^2_{\upsilon k}}{\hat{\sigma}^2_{\upsilon k} + \hat{\sigma}^2_{\varepsilon k}}.$$

Pour conclure, la moyenne μ_d est estimée par

$$\hat{\mu}_d^{\mathrm{BHF}}(t) = \hat{\mu}(t) + \sum_{k=1}^{\mathrm{K}} \hat{\overline{g}}_{dk}\hat{\zeta}_k(t), \quad d = 1,\ldots,\mathrm{D}, \tag{4.13}$$

avec $\hat{\mu}$ et $\hat{\zeta}_k$ les estimations du centre de l'espace et de la k^{eme} composante principale donnés dans les formules (4.7) et (4.8).

Remarque 9. *Plutôt que de projeter les courbes sur les* K *premières composantes principales de l'ACP, on pourrait également choisir d'utiliser d'autres bases de projection, par exemple une base d'ondelettes, celles-ci étant particulièrement adaptées aux courbes irrégulières. Une autre solution consisterait enfin à appliquer les modèles linéaires mixtes fonctionnels sur les valeurs des courbes aux instants de discrétisation ; néanmoins cette façon de faire ne permettrait pas de prendre en compte les corrélations temporelles de la problématique contrairement aux précédentes.*

4.3.2 Régression linéaire fonctionnelle

Les modèles linéaires mixtes que nous venons de présenter peuvent être coûteux en temps de calcul pour de gros volumes de données (c'est-à-dire un nombre K de composantes principales et un nombre D de domaines élevé). Nous proposons donc dans cette sous-section la régression linéaire fonctionnelle, qui est une simplification du modèle précédent ayant comme nous le verrons l'avantage de pouvoir être estimée rapidement même lorsque le nombre d'instants de discrétisation ou de composantes principales considérés est grand. On commence par présenter le modèle puis on propose une méthode rapide pour en estimer les coefficients dans un échantillon en utilisant l'approche citée, entre autres, dans Ardilly (2014).

Estimateurs de courbes moyennes de domaines par régression linéaire fonctionnelle

Dans cette sous-section, on suppose que la première variable explicative de nos vecteurs d'information auxiliaire $\mathbf{X}_i$ est la variable constante égale à 1. Le modèle étudié ici est un cas particulier du modèle général (4.5) dans lequel la fonction de régression f_d définie dans le modèle général peut être définie comme

$$f_d(\mathbf{X}_i, t) = \boldsymbol{\beta}'(t)\mathbf{X}_i, \quad t \in [0, \mathrm{T}]. \tag{4.14}$$

Il s'agit d'une simplification du modèle général dans laquelle les effets des variables explicatives sont linéaires et la fonction de régression f_d ne dépend pas du domaine d. On fait donc l'hypothèse que, conditionnellement aux variables explicatives, la distribution des courbes est la même sur l'ensemble des domaines et qu'il n'y a pas de spécificité des domaines non prise en compte dans les informations auxiliaires.

On est alors dans le contexte usuel de la régression linéaire fonctionnelle (plus précisément dans le cas de la régression d'une variable fonctionnelle sur des variables réelles). Hors du contexte des sondages, ce problème a été étudié notamment par Faraway (1997). Les paramètres de ce modèle peuvent être estimés en projetant les courbes sur une base adaptée (composantes principales ou ondelettes par exemple) ou en estimant la fonction $\boldsymbol{\beta}$ instant par instant.

Si on choisit de travailler dans l'espace des composantes principales, chaque score g_{ik} est modélisé comme suit :

$$g_{ik} = \boldsymbol{\beta}_k'\mathbf{X}_i + \epsilon_{ik}, \quad i \in \mathrm{U}, \quad k = 1, \ldots, \mathrm{K}, \tag{4.15}$$

avec ϵ_{ik} un résidu de moyenne nulle et de variance σ_k. Le paramètre $\boldsymbol{\beta}_k$ de ce modèle est estimé par $\hat{\boldsymbol{\beta}}_k = (\sum_{i\in s}\mathbf{X}_i\mathbf{X}_i')^{-1}\sum_{i\in s}\mathbf{X}_i g_{ik}$, l'estimateur obtenu par les moindres carrés ordinaires (sans pondération) sur l'ensemble de l'échantillon, et on en déduit que le score moyen du domaine d est estimé par :

$$\widehat{\overline{g}}_{dk} = \hat{\boldsymbol{\beta}}_k'\overline{\mathbf{X}}_d, \quad k = 1, \ldots \mathrm{K}.$$

Enfin, l'estimation de la courbe moyenne est obtenue de l'équation (4.13).

Si on ne projette pas les courbes, on a pour tout instant $t \in [0, \mathrm{T}]$, l'estimateur de la courbe de charge moyenne par domaine basé sur l'équation (4.14) :

$$\hat{\mu}_d^{\text{REGLIN}}(t) = \hat{\boldsymbol{\beta}}'(t)\overline{\mathbf{X}}_d, \quad t \in [0, \mathrm{T}], \tag{4.16}$$

avec $\hat{\boldsymbol{\beta}}(t_l)$ le paramètre de la régression linéaire estimé par les moindres carrés ordinaires (sans pondération) sur l'ensemble de l'échantillon :

$$\hat{\boldsymbol{\beta}}(t_l) = \left(\sum_{i \in s} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \sum_{i \in s} \mathbf{X}_i \mathrm{Y}_i(t_l). \tag{4.17}$$

Estimation rapide des coefficients du modèle à l'aide de l'algorithme du calage

L'estimation de ce modèle fonctionnel peut être lourde en temps de calcul si l'on travaille sur de grosses bases de données (beaucoup de domaines ou encore beaucoup d'instants de mesure ou de composantes principales). Cependant, on remarque que notre estimateur par la régression fonctionnelle pour un instant t_l peut s'écrire sous la forme d'une moyenne pondérée des courbes de l'ensemble de l'échantillon :

$$\hat{\mu}_d^{\text{REGLIN}}(t_l) = \frac{1}{\mathrm{N}_d} \sum_{i \in s} w_{id} \mathrm{Y}_i(t_l), \quad l = 1, \ldots, \mathrm{L}, \quad d = 1, \ldots, \mathrm{D}, \tag{4.18}$$

avec des poids w_{id} qui ne dépendent pas de Y ni de temps :

$$w_{id} = \mathbf{X}_i'(\sum_{i \in s} \mathbf{X}_i \mathbf{X}_i')^{-1} \underbrace{\mathrm{N}_d \overline{\mathbf{X}}_d}_{t_{\mathbf{X}_d}}, \quad i \in s. \tag{4.19}$$

Pour implémenter notre estimateur, comme $\mathbf{X}_i$ est constante au cours du temps,il suffit donc, quel que soit le nombre d'instants de discrétisation ou de composantes principales retenues, de déterminer le vecteur des pondérations w_{id} de chacune des unités de l'échantillon.

Pour cela, on procède d'une manière similaire à ce qui est fait dans Ardilly (2014) (bien que cet article relève de l'approche basée sur le plan de sondage alors que nous nous plaçons quant à nous toujours dans le contexte de l'estimation basée sur le modèle) pour déterminer les poids w_{id} en utilisant l'algorithme du calage avec la méthode linéaire.

On remarque en effet que, lorsque le paramètre d'intérêt est une moyenne, notre estimateur par la régression linéaire fonctionnelle défini par l'équation (4.18) et l'estimateur par calage défini par l'équation (4.3) s'écrivent tous deux sous la forme d'une moyenne pondérée des courbes des unités de l'échantillon.

Or, comme montré dans Deville and Särndal (1992), pour la méthode linéaire, l'estimateur par calage est égal à l'estimateur par la régression généralisée qui est un autre estimateur basé sur le plan de sondage, lui-même égal à notre estimateur par régression fonctionnelle pour des poids constants. En effet, l'expression de l'estimateur par régression généralisée (voir Särndal (1992)) de la courbe moyenne du domaine d à l'instant t_l est :

$$\hat{\mu}_d^{\text{GREG}}(t_l) = \frac{1}{\mathrm{N}_d} \left(\sum_{i \in s} d_i \mathrm{Y}_i(t_l) + (t_{\mathbf{X}_d} - \sum_{i \in s} d_i \mathbf{X}_i)' \hat{\boldsymbol{\beta}}_\pi(t_l) \right), \tag{4.20}$$

avec $\hat{\boldsymbol{\beta}}_{\pi}(t_l) = (\sum_{i\in s} d_i \mathbf{X}_i \mathbf{X}_i')^{-1} \sum_{i\in s} d_i \mathbf{X}_i \mathrm{Y}_i(t_l)$ l'estimateur des moindes carrés pondérés du coefficient $\boldsymbol{\beta}(t_l)$ (voir Särndal (1992) ou Rao and Molina (2015)) estimé sur l'ensemble de l'échantillon s.

Donc, si nous appliquons l'algorithme du calage pour la méthode linéaire sur l'ensemble des unités de l'échantillon s en remplaçant les poids de sondage d_i par les poids constants $d_i^* = \frac{\mathrm{N}_d}{n}$, pour tout $i \in s$ en utilisant comme totaux de calage $t_{\mathbf{X}_d}$ le vecteur des totaux des variables auxiliaires sur le domaine d, on calcule l'estimateur $\hat{\boldsymbol{\beta}}_{\pi^*}(t_l) = (\sum_{i\in s} d_i^* \mathbf{X}_i \mathbf{X}_i)^{-1} \sum_{i\in s} d_i^* \mathbf{X}_i \mathrm{Y}_i(t_l)$, et

$$\hat{\mu}_d^{\mathrm{GREG}*}(t_l) = \frac{1}{\mathrm{N}_d}\left(\sum_{i\in s} d_i^* \mathrm{Y}_i(t_l) + (t_{\mathbf{X}_d} - \sum_{i\in s} d_i^* \mathbf{X}_i)' \hat{\boldsymbol{\beta}}_{\pi^*}(t_l)\right). \tag{4.21}$$

Or on a $\hat{\boldsymbol{\beta}}_{\pi^*}(t_l) = \hat{\boldsymbol{\beta}}$ et $\sum_{i\in s} d_i^* \mathrm{Y}_i(t_l) - (\sum_{i\in s} d_i^* \mathbf{X}_i)' \hat{\boldsymbol{\beta}}_{\pi^*}(t_l) = 0$ d'où

$$\begin{aligned}\hat{\mu}_d^{\mathrm{GREG}*}(t_l) &= \frac{1}{\mathrm{N}_d} t_{\mathbf{X}_d}' \hat{\boldsymbol{\beta}}_{\pi^*}(t_l)\\ &= \overline{\mathbf{X}}_d' \hat{\boldsymbol{\beta}}(t_l)\\ &= \hat{\mu}_d^{\mathrm{REGLIN}}(t_l).\end{aligned}$$

Les poids w_{id} fournis par l'algorithme de calage pour la méthode linéaire en partant des poids d_i^* sont donc bien les poids utilisés dans l'équation (4.19). L'utilisation de cet algorithme permet donc d'implémenter notre estimateur de la régression linéaire de manière rapide quel que soit le nombre d'instants de discrétisation en calculant un unique vecteur de poids puis en en déduisant l'estimateur de la courbe moyenne par l'équation (4.18).

On remarquera bien que notre estimateur par la régression linéaire est un estimateur basé sur le modèle et non basé sur le plan. En effet, il ne s'agit pas d'un estimateur par régression généralisée, puisqu'on ne part pas des poids de Horvitz-Thompson mais de poids constants d_i^* qui ne dépendent pas du plan de sondage.

4.3.3 Agrégation de prédictions par arbres de régression pour des courbes

Dans les deux sous-sections à venir, nous présentons des méthodes non paramétriques. Contrairement aux méthodes linéaires, celles-ci n'imposent plus une forme linéaire à la relation entre informations auxiliaires et variable d'intérêt, ce qui permet plus de souplesse dans la modélisation. En contrepartie, ces méthodes ne permettent pas de capter les spécificités des domaines que reflétaient les effets aléatoires du modèle (4.9).

Approche prédictive pour des domaines et estimation non paramétrique

Les deux dernières approches que nous proposons relèvent de *l'approche prédictive* proposée par Valliant et al. (2000) : il s'agit, pour estimer le total ou la moyenne d'une variable sur une population, d'estimer la valeur de cette variable pour chacune

des unités non échantillonnées puis ensuite d'additionner ces prédictions ainsi que les valeurs de la variable pour les unités échantillonnées afin d'en déduire l'estimateur du total. Plus précisément, l'estimateur de la courbe moyenne du domaine d est donnée par (voir Valliant et al. (2000)),

$$\hat{\mu}_d^{\mathrm{PA}}(t) = \frac{1}{\mathrm{N}_d}\left(\sum_{i\in s_d} \mathrm{Y}_i(t) + \sum_{i\in \mathrm{U}_d - s_d} \hat{\mathrm{Y}}_i(t)\right), \quad d = 1,\ldots,\mathrm{D}, \quad t \in [0,\mathrm{T}]. \tag{4.22}$$

Pour obtenir les prédictions individuelles $\hat{\mathrm{Y}}_i(t)$ nous utilisons des modèles non paramétriques : des arbres de régression adaptés aux données fonctionnelles dans cette sous-section et des forêts aléatoires dans 4.3.4. En effet, les arbres de régression pour données fonctionnelles sont fréquemment utilisés à EDF et sont connus pour donner des résultats satisfaisants sur les courbes de consommation électrique. Par ailleurs, dans la littérature, les arbres de régression ont été adaptés au cadre des sondages par Toth and Eltinge (2011) mais pas dans une optique d'estimation de totaux sur des petits domaines.

Dans cette sous-section et la suivante, on cherche donc à estimer un cas particulier du modèle général (4.5) dans lequel la fonction f ne dépend pas du domaine auquel appartient l'unité i,

$$f_d(\mathbf{X}_i, t) = f(\mathbf{X}_i, t) \quad i \in \mathrm{U}, \quad t \in [0,\mathrm{T}]. \tag{4.23}$$

On suppose donc que, conditionnellement aux variables explicatives, la distribution de Y est la même pour l'ensemble des domaines. Tout comme dans le modèle (4.14) et contrairement aux modèles définis dans les équations (4.9) et (4.13), celui-ci ne permettra donc plus de capter d'éventuelles différences entre les domaines non explicables par les informations auxiliaires. En contrepartie, nous ne faisons ici plus l'hypothèse de linéarité des effets des variables auxiliaires $\mathbf{X}_i$ ce qui permet de traduire plus fidèlement la complexité éventuelle du lien entre informations auxiliaires et variable cible.

Par ailleurs, pour utiliser des arbres ou des forêts aléatoires, on a besoin de disposer des informations auxiliaires $\mathbf{X}_i$ pour chaque individu de la population alors que précédemment nous avions seulement besoin des valeurs moyennes $\overline{\mathbf{X}}_d$ sur chacun des domaines de la population et des $\mathbf{X}_i$ sur l'échantillon.

Dans les paragraphes suivants, nous donnons rapidement quelques éléments de bibliographie sur les arbres de régression, en particulier lorsque la variable cible est une courbe. Dans l'Annexe B, nous abordons spécifiquement des points particuliers d'implémentation propres à la problématique de l'estimation de courbes de consommation électrique.

Arbres de régression pour des données fonctionnelles

L'arbre de régression et de classification (CART) proposé par Breiman et al. (1984) est une technique de statistique non paramétrique très populaire. Son objectif est de prédire la valeur d'une variable cible Y en fonction d'un vecteur des variables explicatives $\mathbf{X}_i = (\mathrm{X}_{1i},\ldots,\mathrm{X}_{ji},\ldots\mathrm{X}_{pi}), \quad i \in s$. Pour cela, on détermine un partitionnement de l'espace des $\mathbf{X}_i$ en séparant en deux itérativement le jeu de données, selon une règle de décision (critère de "split") impliquant une unique variable explicative. Cette règle

de décision est choisie parmi toutes les règles possibles de façon à maximiser un critère d'homogénéité (ou, de manière équivalente, minimiser un critère d'inertie) sur chacun des groupes ainsi créés. Ainsi, notre échantillon s constitue le premier nœud λ d'un arbre (sa "racine") que l'on cherche à subdiviser en deux nœuds disjoints λ_l et λ_r tels que $\lambda_l \cup \lambda_r = \lambda$ et $\lambda_l \cap \lambda_r = \emptyset$ de façon à ce que les valeurs de la variable cible Y_i soit les plus homogènes possible dans chacun des nœuds.

Pour les variables X_j quantitatives, les règles de décision sont de la forme

$$\begin{cases} i \in \lambda_l & \text{si} \quad X_{ji} < c \\ i \in \lambda_r & \text{sinon,} \end{cases} \tag{4.24}$$

avec c un point de coupure à optimiser parmi l'ensemble des valeurs possibles de X_j. Pour les variables qualitatives, elles consistent en un découpage en deux sous-ensembles disjoints de modalités.

Le critère d'inertie utilisé pour quantifier l'homogénéité d'un nœud est fréquemment la somme des carrés des résidus c'est-à-dire la somme des carrés des différences entre les valeurs de Y_i pour les unités i du nœud et la moyenne de ces valeurs dans le nœud. Ainsi, pour un nœud λ, soit κ un critère d'inertie, par exemple la somme des carrés des écarts à la moyenne $\kappa(\lambda) = \sum_{i\in\lambda}(Y_i - \overline{Y}_\lambda)^2$ où $\overline{Y}_\lambda$ est la moyenne des Y_i dans le nœud λ. La recherche du critère de split optimal revient à résoudre le problème d'optimisation

$$\underset{\lambda_l,\lambda_r}{\arg\max} \Big(\kappa(\lambda) - \kappa(\lambda_l) - \kappa(\lambda_r)\Big). \tag{4.25}$$

Chacun de ces nœuds sera ensuite à son tour subdivisé en deux nœuds fils et le processus de partitionnement se poursuit jusqu'à atteindre une taille minimale de nœud, jusqu'à ce que la valeur de la variable cible soit la même pour l'ensemble des unités du nœud, ou encore jusqu'à atteindre une profondeur maximale donnée. La partition finale de l'espace est alors constituée par les nœuds finaux de l'arbre, aussi appelés des feuilles. Un résumé de chacune de ces feuilles (très souvent la moyenne pour une variable cible quantitative) devient alors la variable prédite pour l'ensemble des unités affectées à la feuille. Les différents paramètres (taille minimale de nœud et profondeur) peuvent être choisis par validation croisée.

Lorsque la variable Y à prédire n'est plus une variable réelle mais un vecteur de dimension $m > 1$, le principe de l'arbre de régression s'étend très naturellement : l'algorithme de construction de l'arbre et de choix des paramètres par validation croisée reste inchangé mais le critère d'inertie est modifié. Ainsi le critère κ, qui était une distance en dimension 1, est remplacée par une distance en dimension m. Le problème de minimisation s'écrit toujours sous la forme (4.25) mais cette fois le critère est de la forme $\kappa(\lambda) = \sum_{i\in\lambda} ||Y_i - \overline{Y}_\lambda||^2$, où $||.||$ est une distance, par exemple la distance euclidienne ou la distance de Mahalanobis. Les arbres de régression multivariés ont été utilisés par exemple par De'Ath (2002) dans le cadre d'une application à l'écologie.

Enfin, lorsque la variable à prédire Y est une courbe, l'algorithme de construction de l'arbre et de choix des paramètres est identique mais cette fois, on doit utiliser un critère d'inertie κ fonctionnel. De nombreux choix sont possibles. Nous avons choisi de suivre l'approche dite du "Courbotree", décrite dans Stéphan and Cogordan (2009) et

fréquemment employée à EDF pour construire des segmentations de jeux de données de courbes de consommation électrique en fonction de variables explicatives. Dans cette approche, on applique la méthode présentée dans le paragraphe précédent pour Y multivariée sur les vecteurs $\mathbf{Y}_i = (Y_i(t_1), \ldots, Y_i(t_L))$ des valeurs des courbes aux instants de discrétisation, avec la distance euclidienne. La distance euclidienne sur les instants de discrétisation peut alors être vue comme une approximation de la norme $L^2[0,T]$. Plus formellement, le critère fonctionnel s'écrit alors

$$\kappa(\lambda) = \sum_{i \in \lambda} \sum_{l=1}^{L} (Y_i(t_l) - \overline{Y}_\lambda(t_l))^2, \tag{4.26}$$

avec $\overline{Y}_\lambda(t_l) = \frac{\sum_{i \in \lambda} Y_i(t_l)}{n_\lambda}$ où n_λ est le nombre d'unités de l'échantillon appartenant au nœud λ.

Il est d'usage d'élaguer les arbres de régression pour éviter le surapprentissage. Cela n'est toutefois pas encore implémenté dans notre outil Courbotree, et on recommande donc de jouer avec les paramètres de profondeur de l'arbre et de taille minimale des feuilles pour limiter ce surapprentissage. En outre, il n'est pas forcément pertinent d'utiliser les méthodes standard de validation croisée pour déterminer ces paramètres : en effet, celles-ci ont été construites de façon à maximiser la précision de l'estimation de chaque courbe et non pas de la courbe moyenne d'un ensemble d'unités, or il n'est pas assuré que les paramètres optimaux pour l'estimation de la courbe de chaque unité soient également optimaux pour estimer la courbe moyenne d'un domaine.

En pratique, lorsque l'on travaille sur des données de consommation électrique, les courbes considérées ont des niveaux extrêmement hétérogènes, et l'algorithme du Courbotree basé sur la distance euclidienne peut mal fonctionner lorsqu'il est appliqué sur les données brutes. Il est donc fréquent que l'on applique les arbres de régression Courbotree ou les forêts aléatoires CourboForest que nous présentons ci-dessous, sur les formes des courbes, obtenues en divisant celles-ci par leur moyenne. On présente donc en Annexe A une variante de la démarche que nous venons d'exposer qui contexte à séparer la prédiction du niveau des courbes de la prédiction de leur forme. Cette variante est fréquemment utilisée en pratique par les ingénieurs de EDF dans le contexte des courbes de consommation électrique.

4.3.4 Agrégation de prédictions par forêts aléatoires pour des courbes

La littérature met souvent en évidence les médiocres performances prédictives des arbres de régression en comparaison d'autres techniques telles que les SVM (voir par exemple Cristianini and Shawe-Taylor (2000)). En effet, les arbres de régression peuvent être instables et très dépendants de l'échantillon sur lequel ils ont été construits. Pour remédier à ce défaut, Breiman (2001) a proposé l'algorithme des forêts aléatoires. Il s'agit d'une technique ensembliste qui, comme son nom l'indique, consiste à agréger les prédictions issues de différents arbres de régression. Le fait que l'agrégation de prédicteurs instables induise une réduction de variance a été montré notamment dans Breiman et al. (1998). Pour une variable cible quantitative, l'agrégation des prédictions est réalisée en prenant la moyenne des prédictions de chacun des arbres.

Afin de diminuer la variance de la prédiction agrégée, l'objectif est de construire des arbres très différents les uns des autres. L'algorithme de Breiman introduit de la variabilité dans la construction des arbres d'une part en réalisant un rééchantillonnage (tirage aléatoire simple avec remise) des unités et d'autre part en sélectionnant aléatoirement, pour chaque "split" de l'arbre, un sous-ensemble de variables explicatives candidates. Par rapport à un arbre de régression, il y a donc deux paramètres supplémentaires à ajuster pour une forêt aléatoire : le nombre d'arbres et le nombre de variables explicatives candidates à chaque split.

Lorsque la variable d'intérêt est multivariée (ou fonctionnelle), l'algorithme proposé par Breiman s'adapte aisément, en agrégeant les arbres de régression multivariés (ou fonctionnels) présentés dans le paragraphe précédent. Les forêts aléatoires multivariées ont par exemple été étudiées par Segal and Xiao (2011).

L'algorithme que nous proposons ici, appelé "CourboForest", consiste simplement à agréger des arbres de régression fonctionnels construits selon l'approche "Courbotree", c'est-à-dire des arbres de régression multivariés appliqués sur les vecteurs $(\mathbf{Y}_i = \mathrm{Y}_i(t_1), \ldots, \mathrm{Y}_i(t_\mathrm{L}))$ des valeurs des courbes aux instants de discrétisation, avec pour critère de split l'inertie basée sur la distance euclidienne définie par l'équation (4.26).

Pour conclure, chacune des méthodes que nous venons de présenter possède ses avantages et ses inconvénients :

— les modèles linéaires mixtes au niveau unité sont les seuls qui permettent, grâce aux effets aléatoires, d'intégrer dans la modélisation des particularités des domaines non reflétées par les informations auxiliaires.

— la régression linéaire fonctionnelle a la bonne propriété de pouvoir être appliquée en un temps de calcul constant quelle que soit la longueur de la courbe traitée, ce qui peut être intéressant si les volumes de données à traiter sont importants.

— Enfin, les deux méthodes non paramétriques permettent de mieux modéliser des relations non linéaires entre les variables explicatives et la variable d'intérêt. Le choix entre les arbres de régression et les forêts aléatoires dépendra des performances prédictives de ces méthodes sur les données, pour des courbes moyennes de domaines. [2]

Le choix entre ces différentes méthodes dépendra donc des caractéristiques du problème traité et plus précisément de l'importance des différences entre domaines non expliquées par les informations auxiliaires, du caractère non linéaire de la relation entre variables explicatives et variable d'intérêt mais aussi du volume de données à traiter. On présente dans la section 4.5 un cas d'application sur un jeu de données réelles.

2. Il est en effet tout à fait possible que la meilleure des deux méthodes pour la prédiction de chaque courbe individuelle ne soit pas celle qui donne les meilleurs résultats à la maille des domaines.

4.4 Estimation de courbes moyennes robuste aux unités influentes pour des petits domaines

Dans cette section, nous proposons des estimateurs de courbes moyennes pour des petits domaines robustes aux unités influentes. Pour cela, dans la sous-section 4.4.1, nous transformons chacun des estimateurs présentés dans la section précédente en suivant la démarche de construction d'estimateurs robustes en sondages pour des données fonctionnelles présentés dans le Chapitre précédent. Il s'agit donc pour chacune de ces méthodes, d'estimer le biais conditionnel de chaque unité échantillonnée pour l'estimateur de la courbe moyenne de chaque domaine puis de faire apparaître ces biais conditionnels dans l'expression de l'estimateur non robuste afin ensuite de les borner par l'utilisation d'une fonction de Huber (fonctionnelle). On notera toutefois que, pour les méthodes présentées dans la section 4.3, contrairement au Chapitre précédent, on se place dans l'approche des sondages basée sur un modèle et non pas dans l'approche basée sur le plan. La valeur des variables d'intérêt Y n'est donc plus considérée comme déterministe mais comme aléatoire et dans les biais conditionnels les espérances sont donc prises par rapport au modèle de superpopulation et non plus par rapport au plan de sondage.

On propose une expression de biais conditionnel pour chacun des modèles exposés plus haut, d'abord pour les estimateurs reposant sur l'approche basée sur le plan et ensuite pour ceux reposant sur l'approche basée sur un modèle. Nous nous sommes pour cela inspirés de différents travaux existants dans la littérature, hors du contexte des sondages ou hors du contexte des données fonctionnelles.

Ensuite dans la sous-section 4.4.2, on propose une approche alternative consistant à étendre la méthode d'estimation robuste de totaux de variables réelles pour des sous-populations proposée par Sinha and Rao (2009) au cadre des données fonctionnelles. Pour cela, on procède comme précédemment en projetant nos courbes dans un espace de dimension finie puis en appliquant les méthodes préexistantes d'estimation robuste de totaux de variables réelles pour des petites domaines afin d'estimer les totaux des coordonnées des domaines sur chacun des vecteurs de l'espace de projection.

4.4.1 Approches basées sur le biais conditionnel

Tout comme dans le Chapitre 3, nous cherchons à rendre robustes nos estimateurs de courbes moyennes en faisant apparaître les expressions des biais conditionnels de chaque unité dans l'expression des estimateurs non robustes puis en les bornant par une fonction de Huber (fonctionnelle). L'expression de ces biais conditionnels dépend de l'estimateur considéré. D'abord, on traite le cas des estimateurs de Horvitz-Thompson et par calage. Ensuite, on définit la notion de biais conditionnel dans le cadre de l'approche basée sur un modèle et enfin on détaille les expressions de ces biais conditionnels et de leurs estimateurs pour chacun des estimateurs présentés dans la section 4.3.

Approches basées sur le biais conditionnel, pour les méthodes basées sur le plan

Les biais conditionnels de l'estimateur de Horvitz-Thompson et des estimateurs par calage d'une courbe totale ainsi que leurs estimateurs ont déjà été étudiés au Chapitre précédent (équations (3.3) à (3.9) de la section 3.3). Ici, on considère des estimateurs de courbes moyennes et non totales, on doit donc diviser les différentes expressions par les tailles des domaines $\mathrm{N}_d, \quad d = 1, \ldots \mathrm{D}$. Par ailleurs, les estimateurs de Horvitz-Thompson et par calage sont des estimateurs directs, c'est-à-dire que, pour estimer la courbe moyenne d'un domaine d, seules les unités de s_d, l'intersection du domaine d et de l'échantillon s, sont utilisées. Par conséquent les biais conditionnel des unités extérieures au domaine sont donc nuls.

Pour un plan de sondage stratifié, notons $\mathrm{U}_{dh} = \mathrm{U}_h \cap \mathrm{U}_d$ l'intersection du domaine $d = 1, \ldots, \mathrm{D}$ et de la strate $h = 1, \ldots, \mathrm{H}$ dans la population U, de taille N_{dh} et $s_{dh} = s_h \cap \mathrm{U}_d$ l'intersection du domaine $d = 1, \ldots, \mathrm{D}$ et de la strate $h = 1, \ldots, \mathrm{H}$ dans l'échantillon s, de taille n_{dh}. Le biais conditionnel d'une unité $i \in s$ pour l'estimateur de Horvitz-Thompson de la courbe moyenne du domaine $d = 1, \ldots \mathrm{D}$ à l'instant $t \in [0, \mathrm{T}]$ est alors

$$\mathrm{B}_{1i}^{\hat{\mu}_d^{\mathrm{HT}}}(t) = \begin{cases} \frac{1}{\mathrm{N}_d} \frac{\mathrm{N}_{dh}}{\mathrm{N}_{dh}-1} \left(\frac{\mathrm{N}_{dh}}{n_{dh}} - 1 \right) (\mathrm{Y}_i(t) - \overline{\mathrm{Y}}_{dh}(t)) & \text{si} \quad i \in s_d, \\ 0 & \text{si} \quad i \in s - s_d. \end{cases} \tag{4.27}$$

avec $\overline{\mathrm{Y}}_{dh}$ est la moyenne de Y sur U_{dh}.

Pour $i \in s_d$, peut être estimé par :

$$\hat{\mathrm{B}}_{1i}^{\hat{\mu}_d^{\mathrm{HT}}}(t) = \frac{1}{\mathrm{N}_d} \frac{\mathrm{N}_{dh}}{\mathrm{N}_{dh}-1} \left(\frac{\mathrm{N}_{dh}}{n_{dh}} - 1 \right) (\mathrm{Y}_i(t) - \hat{\overline{\mathrm{Y}}}_{dh}(t)), \quad t \in [0, \mathrm{T}], \quad d = 1, \ldots \mathrm{D}, \tag{4.28}$$

avec $\hat{\overline{\mathrm{Y}}}_{dh}$ la moyenne de Y dans s_{dh}.

Pour l'estimateur par calage de la courbe moyenne (défini par les équations (4.3) et (4.4)) du domaine $d = 1, \ldots, \mathrm{D}$ à l'instant $t \in [0, \mathrm{T}]$ et la distance du chi-deux, le biais conditionnel d'une unité échantillonnée $i \in s$ devient

$$\mathrm{B}_{1i}^{\hat{\mu}_d^{cal}}(t) = \begin{cases} \frac{1}{\mathrm{N}_d} \frac{\mathrm{N}_{dh}}{\mathrm{N}_{dh}-1} \left(\frac{\mathrm{N}_{dh}}{n_{dh}} - 1 \right) \left(\mathrm{E}_i(t) - \overline{\mathrm{E}}_{dh}(t) \right), & \text{si} \quad i \in s_d, \\ 0 & \text{si} \quad i \in s - s_d. \end{cases} \tag{4.29}$$

avec E_{di} les résidus de la régression linéaire fonctionnelle (voir 2.1.4) des Y_i sur les variables de calage $\mathbf{X}_i$ sur l'intersection de l'échantillon et du domaine s_d.

Pour $i \in s_d$, il peut être estimé par :

$$\mathrm{B}_{1i}^{\hat{\mu}_d^{cal}}(t) = \frac{1}{\mathrm{N}_d} \frac{\mathrm{N}_{dh}}{\mathrm{N}_{dh}-1} \left(\frac{\mathrm{N}_{dh}}{n_{dh}} - 1 \right) \hat{\mathrm{E}}_i(t), \quad t \in [0, \mathrm{T}], \quad d = 1, \ldots \mathrm{D}, \tag{4.30}$$

avec $\hat{\mathrm{E}}_i(t) = \mathrm{Y}_i(t) - \hat{\boldsymbol{\beta}}_d(t)\mathbf{X}_i$ et $\hat{\boldsymbol{\beta}}_d(t) = (\sum_{i \in s_d} d_i \mathbf{X}_i \mathbf{X}_i')^{-1} \sum_{i \in s_d} d_i \mathbf{X}_i' \mathrm{Y}_i(t)$ l'estimateur des moindres carrés ordinaires sur s_d.

Une fois que l'on dispose d'estimateurs des biais conditionnels, la construction d'estimateurs robustes se déroule comme au Chapitre précédent en bornant les biais conditionnels par une fonction de Huber ψ. Les estimateurs robustes ont donc une forme similaire à celle de l'équation (3.37) mais ici on s'intéresse aux courbes

moyennes de chaque domaine et non plus de l'ensemble de la population. L'expression des estimateurs de courbes moyennes par domaine devient alors

$$\hat{\mu}_d^{\mathrm{R}}(t) = \hat{\mu}_d(t) - \underbrace{\sum_{i\in s} \hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t) + \sum_{i\in s} \psi_{opt;d}\left(\hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t)\right)}_{\Delta_{opt;d}} \quad t \in [0,\mathrm{T}], \quad d = 1,\ldots \mathrm{D}, \tag{4.31}$$

avec $\hat{\mu}_d = \hat{\mu}_d^{\mathrm{HT}}$ ou $\hat{\mu}_d = \hat{\mu}_d^{cal}$ un des estimateurs non robustes de la courbe moyenne du domaine d présentés dans la section précédente et $\hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t)$ un estimateur de $\mathrm{B}_{1i}^{\hat{\mu}_d}(t)$, le biais conditionnel de l'unité i pour cet estimateur.

Plus généralement, l'équation (4.31) décrit l'expression de l'estimateur robuste en fonction de l'estimateur non robuste de la courbe moyenne d'un domaine. Nous l'utiliserons pour construire l'ensemble des estimateurs robustes basés sur la notion de biais conditionnels, qu'on se place dans l'approche basée sur un modèle ou dans l'approche basée sur le plan.

Dans tous les cas, quel que soit l'estimateur non robuste employé, pour déterminer la fonction de troncature optimale $\psi_{opt;d}$ on utilise une des méthodes de construction d'estimateurs robustes présentées au Chapitre précédent. On peut ainsi déterminer une constante univariée $c_{opt}(t_l)$ pour chacun des instants de discrétisation (voir 3.4.1) ou encore projeter les biais conditionnels dans un espace de dimension finie (voir 3.4.2) ou réaliser une troncature fonctionnelle de ces biais conditionnels (voir 3.4.3). On privilégie cependant les méthodes qui ont donné les meilleurs résultats dans les applications sur des données réelles au Chapitre précédent : projection sur la base de l'ACP sphérique, sur la base des ondelettes ou utilisation des instants de discrétisation.

Par exemple, si on choisit de déterminer la fonction de troncature $\psi_{opt;d}$ en appliquant le critère minimax sur les instants de discrétisation, cette équation (4.31) deviendra

$$\hat{\mu}_d^{\mathrm{R}}(t_l) = \hat{\mu}_d(t_l) - \frac{1}{2}\left(\min_{i\in s} \quad \hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t_l) + \max_{i\in s} \quad \hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t_l)\right), \quad \forall l \in 1,\ldots,\mathrm{L}. \tag{4.32}$$

Dans le cas particulier des estimateurs directs de ce paragraphe, seules les unités appartenant au domaine considéré ont un biais conditionnel non nul, et l'expression (4.31) devient alors :

$$\hat{\mu}_d^{\mathrm{R}}(t) = \hat{\mu}_d(t) - \underbrace{\sum_{i\in s_d} \hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t) + \sum_{i\in s_d} \psi_{opt;d}\left(\hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t)\right)}_{\Delta_{opt;d}} \quad t \in [0,\mathrm{T}], \quad d = 1,\ldots \mathrm{D}. \tag{4.33}$$

Biais conditionnels dans l'approche basée sur un modèle

Lorsque nous utilisons les estimateurs de courbes moyennes de domaines dans l'approche basée sur un modèle que nous avons introduits dans la section 4.3, la démarche de construction d'estimateurs robustes à partir d'estimateurs non robustes reste identique à celle présentée dans la sous-section précédente. Cependant, une différence majeure avec le cadre théorique exploité jusqu'à présent est qu'alors on se place dans l'approche basée sous un modèle. Alors, en suivant Beaumont et al. (2013),

le biais conditionnel d'une unité i pour l'estimateur $\hat{\mu}_d$ de la courbe moyenne d'un domaine d est défini comme

$$B_{1i}^{\hat{\mu}_d} = \mathbb{E}_m\left[\hat{\mu}_d - \mu_d | s, Y_i = y_i\right]. \tag{4.34}$$

Cette expression est différente de celle proposée dans la section 3.3 du Chapitre précédent car les valeurs Y_i mais aussi μ_d deviennent maintenant aléatoires. La conception de l'unité influente est alors un peu modifiée : une unité est considérée comme influente si son exclusion de la population a un effet important sur l'erreur de prédiction. Une unité peut être influente parce qu'elle est atypique (le résidu par rapport au modèle est important) mais aussi parce qu'elle a des valeurs des variables explicatives élevées (il s'agit alors d'un point levier). En revanche, son influence ne peut plus venir du plan de sondage et des poids de Horvitz-Thompson.

L'expression de ce biais conditionnel dépend de l'estimateur considéré et du modèle de superpopulation associé.

En outre on remarque que, contrairement à ce que nous avions constaté pour les méthodes basées sur le plan de sondage, ici les biais conditionnels des unités de l'échantillon étrangères aux domaines peuvent apparaître dans cette expression. En effet, pour les méthodes basées sur un modèle, l'ensemble des unités de l'échantillon contribuent à l'estimation du modèle de superpopulation et peuvent donc avoir un impact sur l'expression de l'estimateur.

Estimation robuste par modèles linéaires mixtes sur une base de projection

Dans ce paragraphe nous proposons une version robuste de l'estimateur par modèle linéaire mixte au niveau unité appliqué aux composantes principales de l'ACP proposé dans le paragraphe 4.3.1. Pour cela, nous allons d'une part remplacer la moyenne et les composantes principales de l'ACP classique par la médiane spatiale et les composantes principales sphériques de façon similaire à ce que nous avons proposé dans la section 3.4.2 du Chapitre précédent. D'autre part, nous allons remplacer les estimateurs des scores moyens par domaines par des estimateurs robustes de ces même scores, obtenus en suivant la démarche univariée proposée par Beaumont et al. (2013), pour les biais conditionnels associés à l'estimation des scores moyens par domaines par des modèles linéaires mixtes au niveau unité.

Ainsi, si on projette les courbes dans l'espace des K premières composantes principales sphériques plutôt que dans l'espace des composantes principales classiques, la courbe moyenne d'un domaine d peut être approximée par

$$\mu_d(t) \simeq m(t) + \sum_{k=1}^{K} \underbrace{\left(\frac{1}{N_d}\sum_{i\in U_d} f_{ik}\right)}_{\overline{f}_{dk}} Z_k(t), \quad d = 1,\ldots,D, \quad t \in [0,T], \tag{4.35}$$

avec Z_k, $k = 1,\ldots K$ les composantes principales de l'ACP sphérique, m la médiane spatiale de cet espace et

$$f_{ik} = < Y_i - m, Z_k > = \int_0^T (Y_i - m)(t) Z_k(t)\, dt, \quad i \in U, \quad k = 1,\ldots,K$$

les scores des unités.

De la même manière que pour l'estimateur non robuste de la sous-section 4.3.1, nous considérons pour chaque composante $k = 1,\dots,\mathrm{K}$ un modèle linéaire mixte au niveau unité sur f_{ik} (Rao and Molina (2015), Chapitre 4) :

$$f_{ik} = \boldsymbol{\beta}_k' \mathbf{X}_i + \nu_{dk} + \epsilon_{ik}, \quad i \in \mathrm{U}_d, \quad k = 1,\dots,\mathrm{K}, \tag{4.36}$$

avec $\boldsymbol{\beta}_k' \mathbf{X}_i$ l'effet fixe des informations auxiliaires, ν_{dk} l'effet aléatoire du domaine d et ϵ_{ik} le résidu de l'unité i. On suppose que les effets aléatoires des domaines sont indépendants, et suivent une loi commune de moyenne 0 et de variance $\sigma^2_{\nu k}$. Les résidus sont également indépendants, distribués selon une loi de moyenne 0 et de variance $\sigma^2_{\epsilon k}$. En outre, les effets aléatoires et les résidus sont également supposés indépendants.

En appliquant la démarche de construction d'estimateurs robustes sur les estimations des scores moyens par domaines dans l'équation (4.13), d'une manière similaire à ce qui a été fait dans 3.4.2 pour l'estimation robuste de courbe moyenne hors du contexte des petits domaines, l'estimateur robuste est finalement défini par

$$\hat{\mu}^{\mathrm{R}}_d(t) = \hat{m}(t) + \sum_{k=1}^{\mathrm{K}} \left(\underbrace{-\sum_{i \in s} \hat{\mathrm{B}}_{1i}^{\hat{\bar{f}}_{dk}} + \psi_{c_{dk}}\left(\hat{\mathrm{B}}_{1i}^{\hat{\bar{f}}_{dk}}\right)}_{\Delta_{opt;dk}} \right) \hat{\mathrm{Z}}_k(t), \tag{4.37}$$

avec $\psi_{c_{dk}}$ une fonction de Huber dont la constante c_{dk} aura été sélectionnée par le critère minimax pour la composante $k = 1,\dots\mathrm{K}$ et le domaine $d = 1,\dots\mathrm{L}$. En pratique, comme dans le Chapitre précédent, on n'estime pas explicitement cette constante, mais on calcule directement

$$\Delta_{opt;dk} = \frac{1}{2}\left(\min_{i \in s} \quad \hat{\mathrm{B}}_{1i}^{\hat{\bar{f}}_{dk}}(t) + \max_{i \in s} \quad \hat{\mathrm{B}}_{1i}^{\hat{\bar{f}}_{dk}}(t) \right). \tag{4.38}$$

Pour construire notre estimateur robuste de courbe moyenne, nous devons donc construire des estimateurs $\hat{\mathrm{B}}_{1i}^{\hat{\bar{f}}_{dk}}, i \in s, \quad d = 1,\dots\mathrm{D}$, des biais conditionnels des unités de l'échantillon pour les estimateurs de scores moyens de chaque composante principale. Pour cela, nous allons nous inspirer des estimateurs de biais conditionnels proposés par Jiongo et al. (2013) pour des modèles linéaires mixtes au niveau unité et des estimateurs de totaux de variables réelles.

On introduit les notations suivantes : $\mathbf{X}_d$ la matrice des variables explicatives des unités de s_d (chaque unité en ligne), $\mathbf{1}_{n_d}$ la matrice-colonne de taille $n_d \times 1$ composée de uns, $\mathbf{I}_{n_d}$ la matrice identité de taille $n_d \times n_d$, et enfin $\mathbf{V}_{d;k} = \sigma^2_{\epsilon k}\mathbf{I}_{n_d} + \sigma^2_{\nu k}\mathbf{1}_{n_d}\mathbf{1}_{n_d}'$.

On reprend les notations introduites dans la sous-section 4.3.1. L'estimateur de la moyenne des scores pour la composante k du domaine d peut s'écrire sous la forme d'une somme pondérée des scores des unités de l'échantillon :

$$\hat{\bar{f}}_{dk} = \mathrm{N}_d^{-1} \sum_{i \in s_d} w_{di;k} f_{ik}, \tag{4.39}$$

où $w_{di;k}$ est le poids de l'unité i dans cet estimateur. En effet, en suivant Jiongo et al.

(2013), eq. (12) avec nos notations, on a

$$w_{di;k} = a_{d;k}\mathbf{X}_d' \mathrm{C}_{d;k}^{(i)} \quad \forall i \in s - s_d \tag{4.40}$$

$$= 1 + a_{d;k}\mathbf{X}_d' \mathrm{C}_{d;k}^{(i)} + (\mathrm{N}_d - n_d)\sigma_{\nu k}^2 \mathbf{1}_{n_d}' \mathbf{C}_{d;k}^{(i)} \quad \forall i \in s_d, \tag{4.41}$$

avec $a_{d;k} = \frac{\sum_{i\in \mathrm{U}_d - s_d} \mathbf{X}_i' - \sigma_{\nu dk}^2 (\mathrm{N}_d - n_d)\mathbf{1}_{n_d}' \mathbf{V}_{d;k}^{-1}\mathbf{X}_d}{\mathrm{D}^{-1}\sum_{d=1}^{\mathrm{D}} \mathbf{X}_d' \mathbf{V}_{d;k}^{-1}\mathbf{X}_d}$, ainsi que $\mathbf{C}_{d;k} = \mathbf{V}_{d;k}^{-1}$. Enfin, $\mathbf{C}_{d;k}^{(i)}$ est la i^{eme} colonne de $\mathbf{C}_{d;k}$ et $f_{ik}, \beta_k, \nu_{dk}, \sigma_{\nu k}$ définis dans l'équation (4.9).

Notons en outre,

$$\mathrm{W}_{dd';k} = \begin{cases} \sum_{i\in s_d} w_{di;k} - \mathrm{N}_d & \text{si} \quad d' = d \\ \sum_{i\in s_{d'}} w_{di;k} & \text{si} \quad d' \neq d. \end{cases} \tag{4.42}$$

Selon Jiongo et al. (2013), eq. (18), les biais conditionnels des différentes unités échantillonnées pour l'estimation des scores moyens des composantes principales $\widehat{\widehat{f}}_{dk}, k = 1, \ldots, \mathrm{K}$ sont :

$$\mathrm{B}_{1i}^{\widehat{\widehat{f}}_{dk}} = \begin{cases} \mathrm{N}_d^{-1}(w_{di;k} - 1)(f_{ik} - \mathbf{X}_i'\boldsymbol{\beta}_k - \nu_{dk}) + \mathrm{N}_d^{-1}\mathrm{W}_{dd;k}\nu_{dk} & \text{si} \quad i \in s_d \\ \mathrm{N}_d^{-1} w_{d'i;k}(\mathrm{Y}_i - \mathrm{X}_i'\boldsymbol{\beta}_k - \nu_{d'k}) + \mathrm{N}_d^{-1}\mathrm{W}_{dd';k}\nu_{d'k} & \text{si} \quad i \in s_{d'}, d' \neq d. \end{cases} \tag{4.43}$$

Les biais conditionnels peuvent être estimés en remplaçant les paramètres inconnus $\boldsymbol{\beta}_k$, ν_{dk}, $\sigma_{\epsilon k}^2$, et $\sigma_{\nu k}^2$ par leurs estimateurs. On peut ensuite réinjecter ces expressions d'estimateurs de biais conditionnels dans l'équation (4.37) pour en déduire les estimateurs robustes souhaités.

Estimation robuste par régression linéaire fonctionnelle

On s'intéresse maintenant à l'estimateur par la régression linéaire fonctionnelle défini par l'équation (4.16). Il peut s'écrire sous la forme d'une somme pondérée des courbes de l'échantillon s selon l'équation (4.18). Le poids w_{id} de chaque unité $i \in s$ de l'échantillon est alors défini par l'équation (4.19).

En suivant par exemple Beaumont et al. (2013), pour un instant de discrétisation $t_l, \quad l = 1, \ldots, \mathrm{L}$, le biais conditionnel d'une unité i de l'échantillon est alors

$$\mathrm{B}_{1i}^{\hat{\mu}_d}(t_l) = \mathrm{N}_d^{-1}(w_{id} - 1)(\mathrm{Y}_i(t_l) - \boldsymbol{\beta}'(t_l)\mathbf{X}_i) \quad \forall i \in s. \tag{4.44}$$

Il peut être estimé en remplaçant le paramètre $\boldsymbol{\beta}'(t_l)$ par son estimateur du maximum de vraisemblance :

$$\hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t_l) = \mathrm{N}_d^{-1}(w_{id} - 1)(\hat{\boldsymbol{\beta}}'(t_l)\mathbf{X}_i), \quad \forall i \in s, \tag{4.45}$$

où $\hat{\boldsymbol{\beta}}'(t_l)$ est défini par l'équation (4.17).

Ces estimateurs de biais conditionnels peuvent nous servir à construire des estimateurs robustes à partir des estimateurs non robustes, par l'équation (4.31). La démarche est évidemment identique si on travaille dans la base des composantes principales ou dans toute autre base de projection.

Estimation robuste par approche prédictive et arbres de régression pour des courbes

Intéressons-nous maintenant aux estimateurs obtenus en agrégeant des prédictions issues d'arbres de régression Courbotree. Ces estimateurs ne nécessitent pas de projection des courbes dans des espaces de dimension finie. Ici encore, l'expression de l'estimateur robuste de courbe moyenne peut se déduire de l'expression de l'estimateur non robuste par l'équation (4.31).

A notre connaissance, il n'existe pas dans la littérature d'expression de biais conditionnels pour des arbres de régression. En revanche, Bar-Hen and Poggi (2016) ont proposé différentes mesures d'influences des unités pour les arbres de régression et de classification (CART) dans une optique d'étude de sensibilité et de robustesse de ces arbres. L'article s'intéresse plus particulièrement au problème de la classification (*i.e.* lorsque la variable cible est qualitative) mais la démarche présentée peut être déclinée dans notre cadre où la variable cible Y est quantitative fonctionnelle.

Pour l'approche par arbre de régression, nous proposons donc de remplacer les biais conditionnels estimés $\hat{\mathrm{B}}_{1i}^{\hat{\mu}_d}(t), \quad i \in s, \quad d = 1, \ldots \mathrm{D}$ par une autre mesure d'influence proposée dans Bar-Hen and Poggi (2016), et basée sur le principe du jackknife. Dans cet article, trois indicateurs sont définis pour mesurer l'impact d'une unité sur la prédiction de la variable cible. Nous nous basons sur le premier, qui évalue l'impact de l'absence d'une unité sur l'ensemble des prédictions. Dans le contexte de l'article, qui est celui de la prédiction d'une indicatrice valant 1 ou 0, on note $\hat{\mathrm{Y}}_i = \mathrm{T}(\mathbf{X}_i)$ la prédiction par l'arbre de régression T de la valeur de Y pour l'unité i d'un échantillon de taille n et dont les variables explicatives sont $\mathbf{X}_i$. L'indicateur d'influence de i s'écrit :

$$\mathrm{I}_1(\mathbf{X}_i) = \sum_{j=1}^{n} \mathbf{1}_{\hat{\mathrm{Y}}_j \neq \hat{\mathrm{Y}}_j^{(-i)}}, \quad i \in s,$$

avec $\hat{\mathrm{Y}}_j^{(-i)} = \mathrm{T}^{(-i)}(\mathbf{X}_j)$ la prédiction de Y_j par l'arbre "jackknife" , c'est-à-dire l'arbre construit en utilisant toutes les unités de l'échantillon sauf l'unité i.

Notons $s^{(-i)}, \quad i \in s$ l'échantillon s privé de l'unité i. Dans notre contexte de prédiction d'une variable fonctionnelle Y, l'impact de l'unité i de l'échantillon sur la prévision $\hat{\mathrm{Y}}_j = \hat{f}(\mathbf{X}_j)$ de la courbe de l'unité j de la population peut se définir naturellement comme :

$$\mathrm{I}_{1i}^{\hat{\mathrm{Y}}_j} = \hat{\mathrm{Y}}_j - \hat{\mathrm{Y}}_j^{(-i)}, \quad i \in s, \quad j \in \mathrm{U}$$

où $\hat{\mathrm{Y}}_j^{(-i)} = \hat{f}^{(-i)}(\mathbf{X}_j)$ est la prédiction de la courbe de $j \in \mathrm{U}$ par l'arbre de régression obtenu sur $s^{(-i)}$ avec le même paramétrage (même taille minimale de feuille et même profondeur) que l'arbre utilisé pour la prédiction. En particulier, si comme évoqué dans l'Annexe A on effectue séparément la prédiction du niveau de la courbe et de sa forme, on appliquera la même démarche pour l'arbre "jackknife".

Cet indicateur reflète l'impact de l'unité i sur l'estimation de la courbe de l'unité j. Or, dans notre contexte, on souhaite estimer l'impact de l'unité i sur l'estimation de la courbe moyenne de chacun des domaines $d = 1, \ldots \mathrm{D}$. On va donc adapter naturellement l'indicateur précédent, en considérant que l'impact de l'unité i est l'écart entre l'estimateur de la courbe moyenne du domaine d obtenue sur l'échantillon et l'estima-

teur de la courbe moyenne du domaine privé de i à partir de l'échantillon $s^{(-i)}$. Notons

$$\hat{\mu}_d^{(-i)} = \sum_{j \in s_d} Y_j + \sum_{j \in U_d - s_d} \hat{Y}_j^{(-i)}$$

l'estimateur de la courbe moyenne de d obtenue sur $s^{(-i)}$. On propose alors d'utiliser la mesure d'influence suivante :

$$I_{1i}^{\hat{\mu}_d} = (\hat{\mu}_d - \mu_d) - (\hat{\mu}_d^{(-i)} - \mu_d^{(-i)}), \quad i \in s, \quad d = 1, \ldots D. \tag{4.46}$$

$$= \frac{1}{N_d} \left(\sum_{j \in U_d - s_d} (\hat{Y}_j - \hat{Y}_j^{(-i)}) \right) \quad i \in s, \quad d = 1, \ldots D. \tag{4.47}$$

Cette mesure quantifie la moyenne des écarts entre les prédictions des unités non échantillonnées du domaine obtenues en construisant l'arbre de régression sur l'ensemble de l'échantillon et celles obtenues sur l'échantillon privé de l'unité i avec le même paramétrage. On remarque que cette expression est identique pour les unités appartenant au domaine d et pour les autres.

Pour transformer notre estimateur non robuste par arbre de régression en estimateur robuste, on suggère donc toujours d'appliquer l'équation (4.31) mais en remplaçant les estimateurs de biais conditionnels $\hat{B}_{1i}^{\hat{\mu}_d}$ par nos nouvelles mesures d'influence $I_{1i}^{\hat{\mu}_d}$.

Le calcul de ces influences, et donc la construction d'estimateurs robustes par cette méthode peuvent être coûteux en temps de calcul lorsque les tailles d'échantillons sont grandes. En effet, il est nécessaire de réestimer un nouvel arbre de régression "jackknife" pour chaque unité de l'échantillon.

Estimation robuste par approche prédictive et forêts aléatoires pour des courbes

Dans ce paragraphe, on se place toujours dans l'approche prédictive mais cette fois on considère que les prédictions de courbes individuelles $\hat{Y}_j = \hat{f}(\mathbf{X}_j), \quad j \in U - s$ sont obtenues par des forêts aléatoires Courboforest présentées dans le paragraphe 4.3.4. Tout comme pour les arbres de régression Courbotree, l'expression des influences de chacune des unités échantillonnées est donnée par l'équation (4.46) mais ici, $\hat{Y}_j^{(-i)}$ et $\hat{\mu}_d^{(-i)}$ représentent respectivement les prévisions de la courbe de l'unité $j \in U$ et de la courbe moyenne du domaine $d = 1, \ldots, D$ par la forêt aléatoire construite sur $s^{(-i)}$.

Pour estimer ces influences, une première solution consiste à appliquer une démarche similaire à celle du paragraphe précédent. Pour un domaine $d = 1, \ldots, D$, il faut, pour chaque unité $i \in s$ de l'échantillon, construire une forêt "jackknife" sur $s^{(-i)}$, avec le même paramétrage que la forêt d'origine afin d'en déduire $\hat{Y}_j^{(-i)}, \quad$ pour $j \in U_d - s_d$ puis $\hat{\mu}_d^{(-i)}$, et enfin l'indicateur d'influence souhaité $I_{1i}^{\hat{\mu}_d}, \quad i \in s$. Cependant, cette méthode peut être extrêmement coûteuse en temps de calcul puisqu'elle nécessite de réestimer une nouvelle forêt aléatoire pour chaque unité de l'échantillon.

On va donc proposer une méthode alternative, plus approximative mais aussi plus rapide, à employer lorsque la taille des échantillons considérés nous dissuade d'utiliser la première méthode. On traitera ici le cas simple où on ne sépare pas l'estimation de la forme et l'estimation du niveau. Le cas, très fréquent en pratique pour des courbes

de consommation électrique, où l'on dissocie estimation de la forme et estimation du niveau, est traité dans l'Annexe B.

On cherche à quantifier l'impact de chaque unité i de l'échantillon sur l'estimation de la courbe moyenne de chaque domaine d. Pour cela, on va chercher à exploiter la variabilité induite par la randomisation du tirage des unités dans la forêt aléatoire. Intuitivement, on va chercher à quantifier l'impact de l'unité i sur l'estimation en modélisant le lien entre l'estimation produite par chacun des arbres de la forêt et le nombre de fois où l'unité a été présente dans l'échantillon sur lequel l'arbre a été construit.

Plus précisément, soit une forêt aléatoire CourboForest composée des arbres $a = 1,\ldots,\mathrm{A}$. Soit n_i^a le nombre de fois où l'unité $i \in s$ est sélectionnée dans l'échantillon utilisé pour construire l'arbre a. Ce nombre peut être supérieur à 1 si le tirage est réalisé avec remplacement. Enfin, on note $\hat{\mathrm{Y}}_j^a$ la courbe prédite pour l'unité $j \in \mathrm{U}$ par l'arbre a. La courbe moyenne estimée sur le domaine d à partir des prédictions de l'arbre de régression a est

$$\hat{\mu}_d^a = \frac{1}{\mathrm{N}_d}\left(\sum_{j\in s_d} \mathrm{Y}_j + \sum_{j\in \mathrm{U}_d - s_d} \hat{\mathrm{Y}}_j^a\right). \tag{4.48}$$

Pour chaque unité i et chaque domaine d, on suppose qu'il existe une relation linéaire entre la prédiction de la courbe moyenne du domaine obtenue par l'arbre a et le nombre n_i^a de sélections de l'unité i dans l'arbre a.

$$\hat{\mu}_d^a = \mathrm{A}_{di} + \mathrm{B}_{di} n_i^a + \mathrm{E}_{di}^a, \quad \forall i \in s, \quad d = 1,\ldots,\mathrm{D}$$

avec E_{di}^a un résidu d'espérance nulle. Intuitivement, le cœfficient B_{di} traduit alors l'influence de l'unité i (présente en un exemplaire) sur la prédiction de la courbe moyenne du domaine d . On prendra donc comme mesure d'influence l'estimateur des moindres carrés ordinaires de ce coefficient :

$$\mathrm{I}_{1i}^{\hat{\mu}_d} = \hat{\mathrm{B}}_{di} = \frac{\sum_{a=1}^{\mathrm{A}}(n_i^a - \sum_{a=1}^{\mathrm{A}} n_i^a)(\hat{\mu}_d^a - \hat{\mu}_d)}{\sum_{a=1}^{\mathrm{A}}(n_i^a - \sum_{a=1}^{\mathrm{A}} n_i^a)}, \quad i \in s, \quad d = 1,\ldots\mathrm{D}.$$

Si le nombre de sélections n_i^a de l'unité i est constant pour l'ensemble des arbres $a = 1,\ldots\mathrm{A}$, on ne peut pas réaliser cette régression et on prend $\mathrm{I}_{1i}^{\hat{\mu}_d} = 0$.

4.4.2 Estimateurs REBLUP sur composantes principales sphériques

Dans cette sous-section, on propose une démarche alternative pour la construction d'estimateurs robustes. On se place dans le même contexte que pour la sous-section 4.4.1 : la relation entre variables explicatives et variable d'intérêt peut être décrite par un modèle linéaire mixte au niveau unité posé sur chacune des composantes principales sphériques (voir équation (4.36)). L'estimateur de la courbe moyenne des domaines se déduit des estimateurs des scores moyens des domaines par l'équation (4.35).

Pour estimer les paramètres $\boldsymbol{\beta}_k$ et ν_{dk} des modèles linéaires mixtes au niveau unité sur les scores de chacune des composantes principales dans (4.36), on ne va plus utiliser les estimateurs du maximum de vraisemblance restreint mais plutôt des estimateurs robustes de la famille des M-estimateurs, proposés par Sinha and Rao (2009) et

appelés REBLUP. Cette approche devra toutefois être considérée avec prudence, car peut générer des biais si la distribution des erreurs n'est pas symétrique : en effet, pour reprendre la classification de Chambers (1986), il s'agit d'une méthode *Robust Projective* et non *Robust Predictive* c'est-à-dire qu'elle considère que les unités non échantillonnées se comportent comme les unités non atypiques de l'échantillon et qu'elle n'intègre pas de correction du biais.

Dans le cas précis de nos modèles linéaires mixtes, ces estimateurs sont obtenus en introduisant des fonctions de Huber dans les équations d'estimation du maximum de vraisemblance utilisées pour estimer les paramètres. Cela permet en effet de diminuer le poids d'éventuelles unités influentes dans ces estimations. Les équations sont ensuite résolues numériquement par un algorithme de type Newton-Raphson de manière itérative pour mener à l'estimation des paramètres recherchés.

Nous utilisons cette méthode pour produire des estimateurs de scores moyens des domaines pour chaque composante principale puis en déduire ensuite les estimateurs robustes de courbes moyennes de domaines par l'expansion de Karhunen-Loève. Tout comme pour l'estimateur non robuste, il est également possible d'utiliser d'autres bases de projection ou de travailler instant par instant sur les courbes discrétisées.

Plus précisément, introduisons quelques notations matricielles, similaires à celles de la sous-section 4.4.1 : soient $\mathbf{X}$ la matrice des variables explicatives des unités de s (chaque unité en ligne), $\mathbf{1}_n$ la matrice-colonne de taille $n \times 1$ composée de uns, et enfin $\mathbf{I}_n$ la matrice identité de taille $n_\times n$.

Pour tout $k = 1, \ldots, \mathrm{K}$, soient $\mathbf{f}_k^{'} = (f_{1k}, \ldots, f_{nk})$ le vecteur des scores des unités de l'échantillon, $\boldsymbol{\nu}_k = (\nu_{1k}, \ldots, \nu_{dk})$ le vecteur des effets aléatoires des domaines et enfin $\mathbf{V}_k = \sigma^2_{\epsilon k}\mathbf{I}_n + \sigma^2_{\nu k}\mathbf{1}_n\mathbf{1}_n^{'}$. Avec nos notations, l'estimateur BLUP (non robuste) $\tilde{\boldsymbol{\beta}}_k$ de $\boldsymbol{\beta}_k$ est donné par (voir Rao and Molina (2015), Chapitre 5) :

$$\tilde{\boldsymbol{\beta}}_k = (\mathbf{X}^{'}\mathbf{V}_k^{-1}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{V}_k^{-1}\mathbf{f}_k \tag{4.49}$$

et l'estimateur BLUP des effets aléatoires est :

$$\tilde{\boldsymbol{\nu}}_k = \underbrace{\sigma^2_{\nu k}\mathbf{1}_n^{'}}_{\boldsymbol{\Sigma}_{\nu k}}\mathbf{V}_k^{-1}(\mathbf{f}_k - \mathbf{X}^{'}\tilde{\boldsymbol{\beta}}_k). \tag{4.50}$$

On montre aisément que les estimateurs (4.49) et (4.50) sont solutions des équations

$$\mathbf{X}^{'}\mathbf{V}_k^{-1}(\mathbf{f}_k - \mathbf{X}^{'}\tilde{\boldsymbol{\beta}}_k) = 0 \tag{4.51}$$

et

$$\boldsymbol{\Sigma}_{\nu k}\mathbf{V}_k^{-1}(\mathbf{f}_k - \mathbf{X}^{'}\tilde{\boldsymbol{\beta}}_k) - \tilde{\boldsymbol{\nu}}_k = 0 \tag{4.52}$$

Une solution évidente pour construire des estimateurs robustes consiste donc à introduire des fonctions de Huber dans ces équations pour limiter l'influence d'éventuels outliers :

$$\mathbf{X}^{'}\mathbf{V}_k^{-1/2}\psi_c\left(\mathbf{V}_k^{-1/2}(\mathbf{f}_k - \mathbf{X}^{'}\tilde{\boldsymbol{\beta}}_k)\right) = 0 \tag{4.53}$$

et

$$\boldsymbol{\Sigma}_{\nu k}\mathbf{V}_k^{-1/2}\psi_c\left(\mathbf{V}_k^{-1/2}(\mathbf{f}_k - \mathbf{X}^{'}\tilde{\boldsymbol{\beta}}_k)\right) - \boldsymbol{\Sigma}_{\nu k}^{1/2}\psi_c\left(\boldsymbol{\Sigma}_{\nu k}^{-1/2}\tilde{\boldsymbol{\nu}}_k\right) = 0 \tag{4.54}$$

avec $c > 0$ le paramètre de troncature à choisir.

Les solutions de ces équations peuvent être numériquement instables, c'est pourquoi Fellner (1986) a suggéré une écriture alternative. Comme les paramètres de variances σ_{ν_k} et σ_{ϵ_k} sont inconnus, il est nécessaire de disposer également d'estimateurs robustes de ces quantités, c'est pourquoi Richardson and Welsh (1995) ont proposé des estimateurs robustes par maximum de vraisemblance restreint de celles-ci. Enfin, pour l'estimation de totaux de variables réelles pour de petits domaines par modèles linéaires mixtes au niveau unité, Sinha et Rao proposent une approche de construction d'estimateurs robustes qui se base sur l'ensemble de ces travaux.

Pour une composante $k = 1, \ldots, \mathrm{K}$, les estimateurs robustes des paramètres du modèle sont finalement obtenus en résolvant numériquement les équations estimantes

$$\mathbf{X}'\mathbf{V}_k^{-1}\mathbf{U}_k^{-1/2}\psi_c\left(\mathbf{U}_k^{-1/2}(\mathbf{f}_k - \mathbf{X}'\boldsymbol{\beta}_k)\right) = 0, \tag{4.55}$$

où $\mathbf{U}_k = diag(\mathbf{V}_k)$ et ψ_c est une fonction de $\mathbb{R}^p \to \mathbb{R}^p$ dont le paramètre de troncature $c > 0$ doit être fixé et

$$\psi_c\left((\mathbf{f}_k - \mathbf{X}'\boldsymbol{\beta}_k)'\mathbf{U}_k^{1/2}\right)\mathbf{U}_k^{1/2}\mathbf{V}_k^{-1}\left(\frac{\partial \mathbf{V}_k}{\partial \epsilon}\right)\mathbf{V}_k^{-1}\mathbf{U}_k^{1/2}\psi_c\left(\mathbf{U}_k^{-1/2}(\mathbf{f}_k - \mathbf{X}'\boldsymbol{\beta}_k)\right) = tr\left(\mathrm{D}_k^c(\frac{\partial \mathbf{V}_k}{\partial \epsilon})\right) \tag{4.56}$$

pour $\epsilon = \{\sigma_{\epsilon k}, \sigma_{\nu k}\}$ et $\mathrm{D}_k^c = \mathbb{E}\left(\psi_c^2(\mathrm{Z})\right)\mathbf{V}_k^{-1}$ où $\mathrm{Z} \sim \mathcal{N}(0,1)$

Nous pouvons ensuite finalement en déduire des estimateurs robustes des scores moyens des domaines $\overline{f}_{dk}$ et ensuite en déduire les estimateurs de courbes moyennes par domaine en utilisant l'équation (4.35).

4.5 Application à des courbes de consommation électrique

Nous allons maintenant tester les méthodes que nous venons de présenter afin de comparer leurs performances sur des données de consommation électrique de clients résidentiels français.

4.5.1 Présentation du jeu de données

Dans ce chapitre, nous avons choisi de travailler sur un jeu de données différent de celui utilisé dans le reste de la thèse. En effet, les courbes irlandaises que nous utilisons par ailleurs ne se prêtaient pas à l'illustration de notre problématique puisque nous n'avions aucune information de nature géographique qui aurait pu nous permettre de découper la population en domaines. De plus, de premiers tests on montré que les variables explicatives fournies en complément des données de consommation électrique n'avaient qu'un pouvoir explicatif très modéré sur les consommations. La qualité de la modélisation aurait donc été médiocre pour l'ensemble des méthodes proposées et il aurait été délicat de comparer les performances des différentes approches.

Nous avons donc travaillé sur un jeu de données appartenant à EDF et qui contient des courbes de consommation électrique de N = 1905 clients résidentiels français au pas journalier d'octobre 2011 à mars 2012, sans valeur manquante (L = 177 points).

Cette population se subdivise en D = 8 domaines correspondant à des zones géographiques.

Nous sommes conscients que la volumétrie de ce jeu de données est très faible, ce qui pourrait remettre en question la pertinence des conclusions de nos tests. Cependant, d'autres tests ont été réalisés sur des jeux de données beaucoup plus importants (centaines de milliers de courbes et dizaines de domaines) selon le même protocole pour d'autres clients résidentiels avec des variables explicatives assez similaires. Malheureusement, pour des raisons de confidentialité, nous ne pouvons pas présenter les résultats de ces tests. Cependant, les conclusions méthodologiques obtenues à partir des deux jeux de données étaient globalement similaires.

Pour chacun des individus de notre population d'étude, on dispose de plusieurs variables auxiliaires au niveau individuel : puissance souscrite, option tarifaire (Base ou Heures Creuses) [3], consommation annuelle de l'année précédente, type de logement (appartement ou maison individuelle) et proportion des consommations annuelles survenant l'été.

L'ensemble des tests ont été implémentés en R. Pour les modèles linéaires mixtes niveau unité, nous avons utilisé le package `rsae` pour le REBLUP de Sinha and Rao (2009) et le package `sampling` pour la régression fonctionnelle utilisant l'algorithme du calage et l'estimateur par calage. L'estimation d'arbres de régression et de forêts aléatoires pour des courbes est effectuée grâce à un package interne EDF, nommé `courbotree`, que nous avons développé, et qui réalise des appels à des modules java préexistants pour la construction des arbres et leur visualisation interactive.

4.5.2 Protocole de test

Nous comparons différents estimateurs obtenus en suivant les méthodes exposées dans ce chapitre, pour différentes modélisations (estimateur de Horvitz-Thompson, estimateur par calage, modèles linéaires mixtes au niveau unité, régressions linéaires fonctionnelles, arbres de régression, forêts aléatoires). Nous testons plusieurs versions du modèle linéaire mixte au niveau unité, en posant les modèles linéaires mixtes sur les scores de l'ACP comme suggéré dans le paragraphe 4.3.1, ou sur les coefficients d'une base d'ondelettes, ou encore en les appliquant directement sur les valeurs de la courbe aux instants de discrétisation. On teste les versions robustes et non robustes de chacun de ces modèles. On teste également l'application des estimateurs de Sinha Rao sur les instants de discrétisation ou sur les composantes principales de l'ACP.

Pour les méthodes non paramétriques, les forêts et les arbres ont une profondeur de 5 et une taille minimale de feuille de 5. Le nombre d'arbres des forêts est de 40. Les algorithmes peuvent être appliqués en séparant l'estimation du niveau de la courbe et de sa forme comme préconisé dans l'Annexe B (normalisation = "oui") ou pas (normalisation = "non").

Le calage est réalisé selon la méthode du raking ratio et, lorsque celui-ci ne

3. dans l'option Base, le prix du kWh reste constant, tandis que dans le tarif Heures Creuses, il est réduit pendant huit heures (dites creuses). Ce dernier tarif a tendance a être privilégié par les plus gros consommateurs. Les horaires d'heures creuses peuvent varier d'un client à l'autre, mais ce facteur n'a pas d'impact ici puisque nous travaillons au pas journalier.

converge pas, par la méthode linéaire. En outre, le nombre d'unités appartenant au plus petit des domaines pouvant être très faible pour certaines simulations, on ne réalisera le calage que pour les domaines d'au moins 10 unités et on utilisera un estimateur de Horvitz-Thompson sinon.

Afin de ne pas multiplier les combinaisons possibles, nous nous sommes finalement concentrés sur les estimateurs énumérés dans la Table 4.1. Le paramétrage des méthodes par arbres de régression ou forêts aléatoires est détaillé dans la Table 4.2.

titre	référence	robuste	projection
Horvitz-Thompson (naïf)	eq. (4.2)	non	aucune
calage	eq. (4.3)	non	aucune
modèle linéaire mixte	section (4.3.1)	non	aucune
modèle linéaire mixte sur acp	section (4.3.1)	non	ACP
modèle linéaire mixte sur ondelettes	section (4.3.1)	non	ondelettes
régression linéaire	section (4.3.2)	non	aucune
courbotree	section (4.3.3)	non	aucune
courbotree normalise	section (4.3.3)	non	aucune
courboforest	section (4.3.3)	non	aucune
Horvitz-Thompson robuste	par. (4.4.1)	oui	aucune
calage robuste	eq. (4.3)	non	aucune
modèle linéaire mixte robuste	par. (4.4.1)	oui	aucune
modèle linéaire mixte robuste sur acp	par. (4.4.1)	oui	ACP
modèle linéaire mixte robuste sur ondelettes	par. (4.4.1)	oui	ondelettes
régression linéaire robuste	par. (4.4.1)	oui	aucune
régression linéaire robuste sur ACP	par. (4.4.1)	oui	aucune
courbotree robuste	par. (4.4.1)	oui	aucune
courboforest robuste	par. (4.4.1)	oui	aucune
sinha rao	section (4.4.2)	oui	aucune
sinha rao sur ACP	section (4.4.2)	oui	ACP

TABLEAU 4.1 – Différentes méthodes d'estimation testées.

titre	profondeur	nombre d'arbres	normalisation
courbotree	5	1	non
courbotree normalise	5	1	oui
courboforest	5	40	non
courbotree robuste	5	1	non
courboforest robuste	5	40	non

TABLEAU 4.2 – Paramétrage des arbres et forêts aléatoires.

Afin d'évaluer la qualité de nos méthodes d'estimation, notre protocole de test consiste à réaliser un grand nombre E de simulations de tirage d'échantillons parmi notre population de départ et ensuite à estimer la courbe moyenne de chacun des D = 8 domaines par les différentes méthodes proposées à partir de chaque échantillon tiré. Nous réalisons E = 1000 simulations. Dans nos simulations, le huitième domaine

sera toujours non échantillonné, afin de mesurer la performance de nos différents estimateurs dans ce cas de figure. Pour chaque simulation, on sélectionne par sondage aléatoire simple $n = 200$ individus parmi ceux appartenant aux $\mathrm{D} - 1 = 7$ domaines échantillonnés.

Des indicateurs de qualité sont ensuite calculés pour chaque domaine en comparant les courbes estimées aux courbes réelles. Nous synthétisons finalement ces résultats en séparant d'une part les performances des méthodes sur les domaines échantillonnés (mesurées par les moyennes des indicateurs sur les sept domaines échantillonnés) et d'autre part leurs performances sur le domaine non échantillonné.

Des indicateurs de qualité sont ensuite calculés pour chaque domaine en comparant les courbes estimées aux courbes réelles. Ces indicateurs sont les mêmes qu'au Chapitre précédent (voir 3.6.2). Nous synthétisons ensuite ces résultats en séparant d'une part les performances des méthodes sur les domaines échantillonnés (mesurées par les moyennes des indicateurs sur les sept domaines échantillonnés) et d'autre part leurs performances sur le domaine vide.

Plus précisément, considérons que le domaine non échantillonné est le domaine $d = \mathrm{D} = 8$ et que les autres sont échantillonnés. Soit $\mu_d(t_l)$ la courbe moyenne du domaine d à l'instant t_l et $\hat{\mu}_d(t_l)$ son estimateur par une méthode donnée. On note $\mathrm{E}_{\mathrm{MC}}[\hat{\mu}_d(t_l)] = \frac{1}{\mathrm{E}}\sum_{e=1}^{\mathrm{E}} \hat{\mu}_d^e(t_l)$ l'espérance Monte-Carlo de l'estimateur $\hat{\mu}_d(t_l)$ où $\hat{\mu}_d^e(t_l)$ est l'estimateur de la courbe moyenne obtenu pour la simulation $e = 1, \ldots, \mathrm{E}$.

Pour un instant $t_l, l = 1, \ldots, \mathrm{L}$ donné, on construit d'abord un indicateur de biais,

$$\mathrm{RB}(\hat{\mu}_d)(t_l) = 100\frac{|\mathrm{E}_{\mathrm{MC}}[\hat{\mu}_d(t_l)] - \mu_d(t_l)|}{\mu_d(t_l)}. \tag{4.57}$$

On définit ensuite l'indicateur d'erreur globale (erreur quadratique moyenne ou MSE),

$$\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d)(t_l) = \frac{1}{\mathrm{E}}\sum_{e=1}^{\mathrm{E}} \left(\hat{\mu}_d^e(t_l) - \mu_d(t_l)\right)^2. \tag{4.58}$$

Plus cet indicateur global sera faible, plus la qualité de l'estimateur sera considérée comme bonne. L'erreur quadratique moyenne peut être difficile à appréhender, on va donc utiliser un troisième indicateur plus facile à lire appelé Efficacité Relative (RE), qui compare l'erreur quadratique moyenne Monte-Carlo $\mathrm{MSE}_{\mathrm{MC}}$ de chaque méthode avec celle d'un estimateur de référence,

$$\mathrm{RE}(\hat{\mu}_d)(t_l) = 100\frac{\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d)(t_l)}{\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d^{\mathrm{HT}})(t_l)}, \quad d \in 1, \ldots, \mathrm{D}-1. \tag{4.59}$$

Plus l'indicateur RE sera faible, plus l'estimateur sera considéré comme performant. Un RE de 100 correspond à un indicateur aussi performant que l'estimateur de référence.

Ici l'estimateur de référence $\hat{\mu}_{\mathrm{Y}d}^0 = \hat{\mu}_d^{\mathrm{HT}}$ est l'estimateur de Horvitz-Thompson (qui pour notre plan de sondage aléatoire simple est la moyenne simple des courbes du domaine considéré), il correspond au modèle décrit par l'équation (4.2) et sera aussi désigné par le terme "naïf". Pour le domaine non échantillonné, cet estimateur ne peut pas

être calculé. L'estimateur RE est alors obtenu en divisant les MSE des différents estimateurs par le MSE moyen de l'estimateur de Horvitz-Thompson sur les sept domaines échantillonnés, *i.e.*,

$$\mathrm{RE}(\hat{\mu}_{\mathrm{D}})(t_l) = 100\frac{\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_{\mathrm{D}})(t_l)}{\overline{\mathrm{MSE}}^{\mathrm{HT}}_{\mathrm{MC}}(t_l)}, \quad l = 1,\ldots,\mathrm{L}, \tag{4.60}$$

avec

$$\overline{\mathrm{MSE}}^{\mathrm{HT}}_{\mathrm{MC}}(t_l) = \sum_{d=1}^{\mathrm{D}-1} \mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}^{\mathrm{HT}}_d)(t_l).$$

Pour chaque indicateur et chaque instant $t_l, l = 1,\ldots,\mathrm{L}$, les résultats obtenus sur les différents domaines échantillonnés sont ensuite agrégés :

$$\mathrm{RB}_{ech}(\hat{\mu})(t_l) = \frac{1}{\mathrm{D}-1}\sum_{d=1}^{\mathrm{D}-1} \mathrm{RB}(\hat{\mu}_d)(t_l). \tag{4.61}$$

$$\mathrm{MSE}_{ech}(\hat{\mu})(t_l) = \frac{1}{\mathrm{D}-1}\sum_{d=1}^{\mathrm{D}-1} \mathrm{MSE}(\hat{\mu}_d)(t_l). \tag{4.62}$$

$$\mathrm{RE}_{ech}(\hat{\mu})(t_l) = \frac{1}{\mathrm{D}-1}\sum_{d=1}^{\mathrm{D}-1} \mathrm{RE}(\hat{\mu}_d)(t_l). \tag{4.63}$$

Par contre, les indicateurs obtenus sur le domaine non échantillonné sont utilisés tels quels.

Afin d'évaluer la performance globale, on considère finalement la moyenne de ces indicateurs sur l'ensemble des instants de la période de test, en séparant toujours les domaines échantillonnés du domaine non échantillonné. On s'intéresse également au temps de calcul des différents estimateurs.

4.5.3 Résultats et conclusion des tests

Les résultats des tests des méthodes sont présentés dans les Tables 4.3 et 4.4 et illustrés par les Figures 4.1 à 4.5.

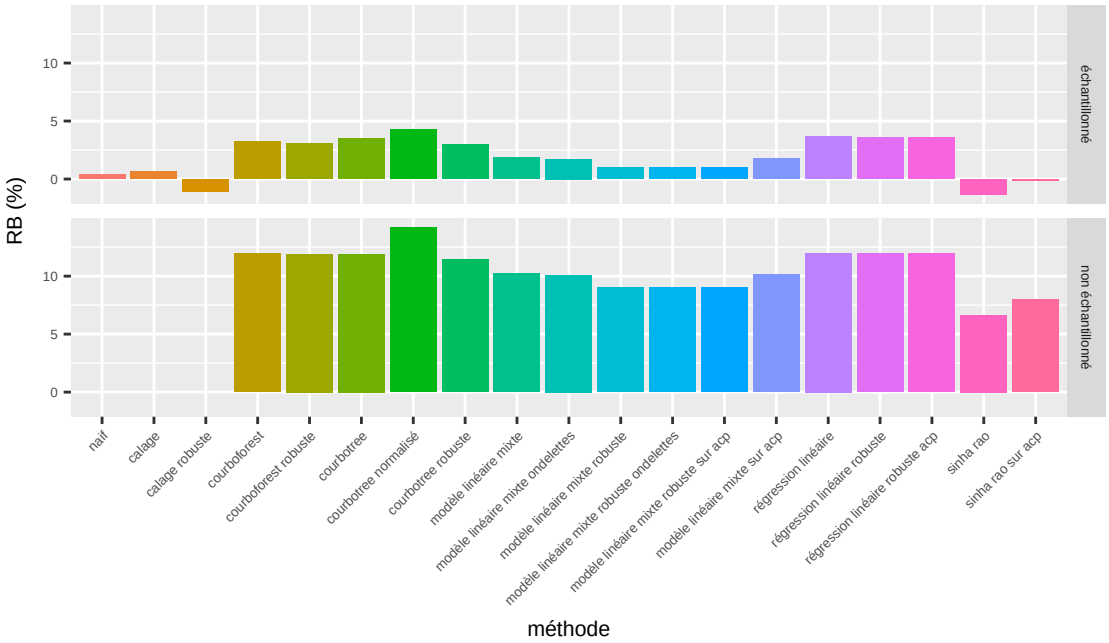

FIGURE 4.1 – Moyennes des biais relatifs en % (formule (4.57)) des méthodes d'estimation, sur l'ensemble des instants et des domaines, en séparant les domaines échantillonnés et non échantillonnés.

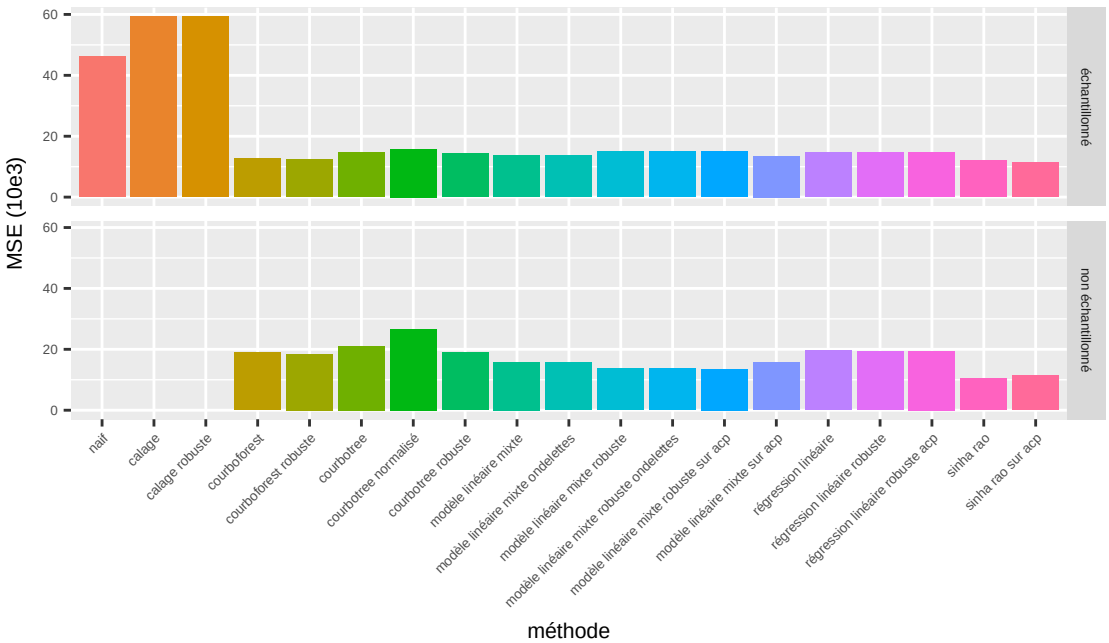

FIGURE 4.2 – Moyennes des MSE (formule (4.58)) des méthodes d'estimation, sur l'ensemble des instants et des domaines, en séparant les domaines échantillonnés et non échantillonnés.

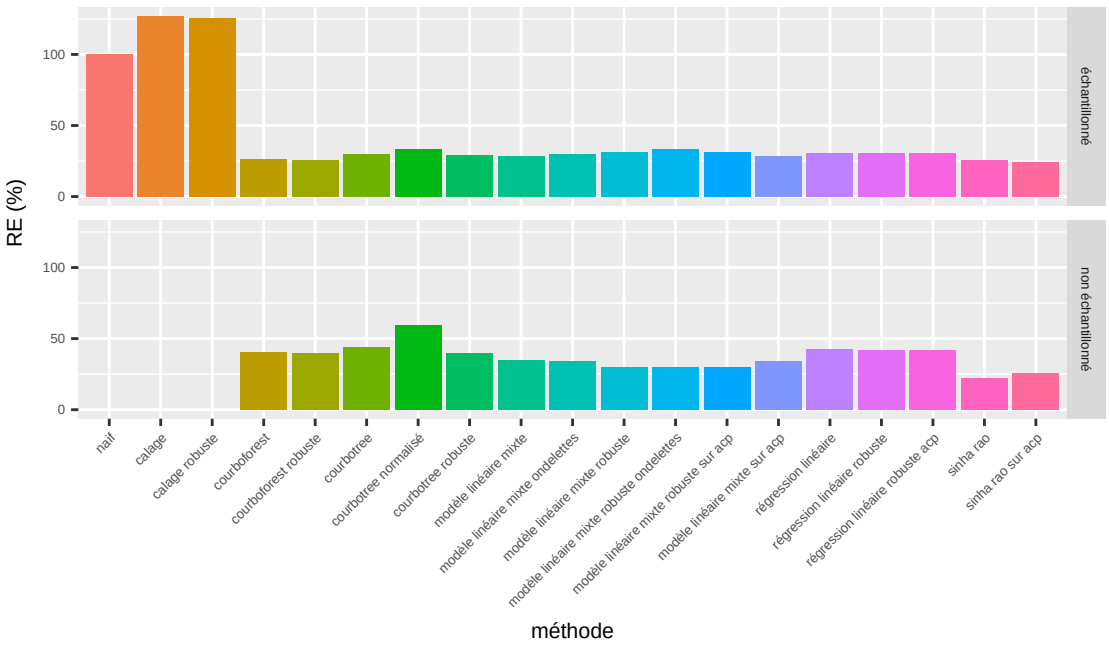

FIGURE 4.3 – Moyennes des efficacité relatives (RE, en %) (formule (4.59)) des estimateurs, sur l'ensemble des instants et des domaines, en séparant les domaines échantillonnés et non échantillonnés.

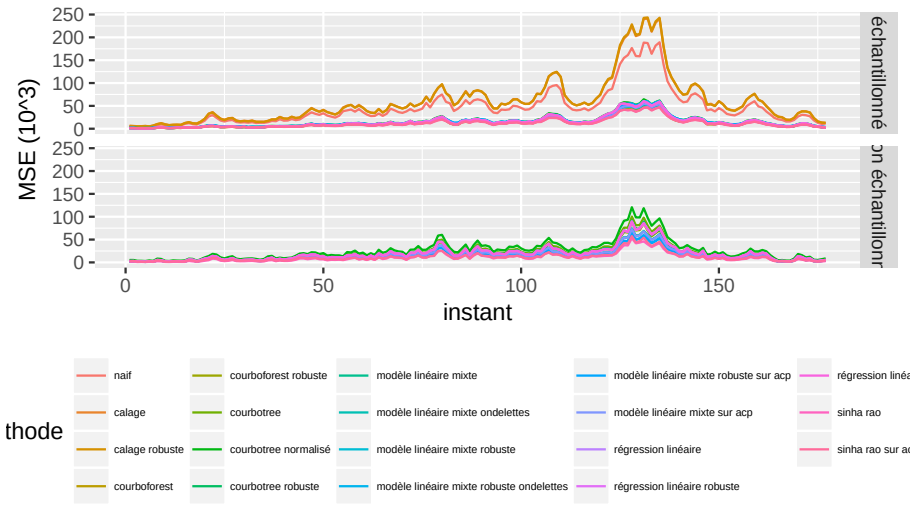

FIGURE 4.4 – Évolution de la moyenne des MSE des différents estimateurs au cours du temps, en séparant les domaines échantillonnés et non échantillonnés.

On remarque que sur cet exemple, l'estimateur direct par calage donne de moins bons résultats que l'estimateur de Horvitz-Thompson. Cela s'explique par le fait que les effectifs peuvent être faibles pour le plus petit des domaines, ce qui rend l'estimateur par calage potentiellement très instable (même si on ne réalise pas de calage lorsque les domaines contiennent moins de 10 individus dans l'échantillon). Pour les

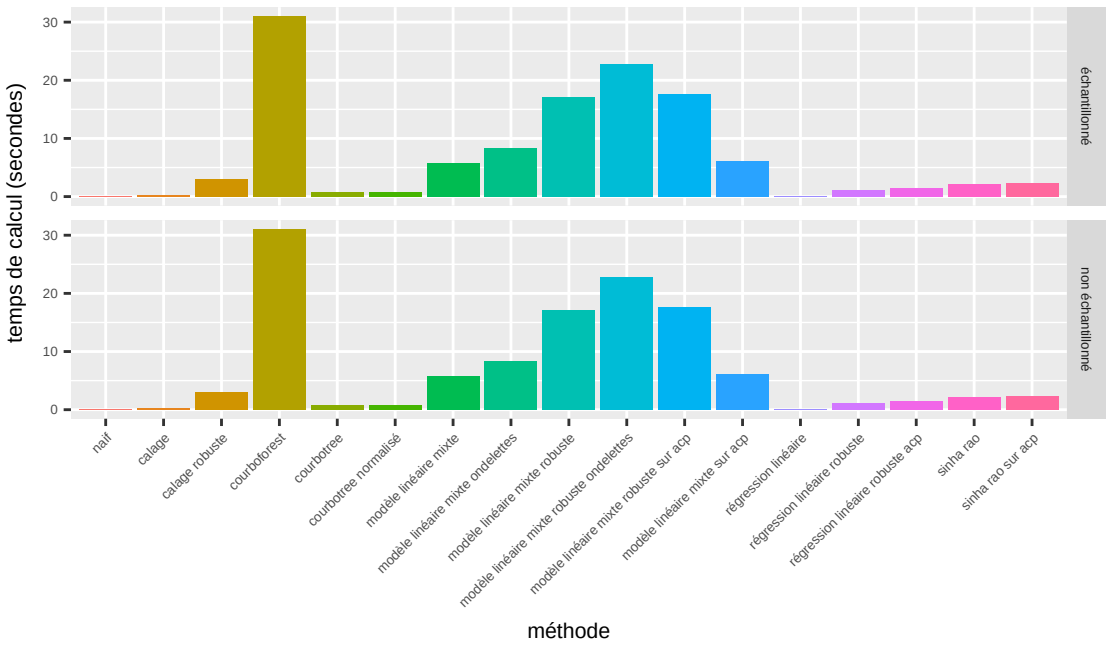

FIGURE 4.5 – Temps de calculs moyens en secondes des différents estimateurs hors courbotree robuste et courboforest robuste pour un échantillon de 200 courbes de 177 points

autres méthodes, on constate que le fait d'intégrer des variables explicatives dans l'estimation induit un net gain de performances par rapport à l'estimateur de Horvitz-Thompson : ainsi, les efficacités relatives se situent autour de 30% sur les domaines échantillonnés pour la plupart des méthodes non robustes.

Que ce soit sur les domaines échantillonnés ou non échantillonnés, les meilleures méthodes sont celles qui reposent sur l'estimateur de Sinha Rao. Elles donnent des résultats sensiblement meilleurs à ceux des modèles linéaires mixtes non robustes.

Sur les domaines échantillonnés, la méthode Courboforest et sa version robuste arrivent en seconde position suivies par les modèles linéaires mixtes, le Courbotree et les régressions linéaires, relativement proches les unes des autres. On remarque que les méthodes non paramétriques ont un biais relatif légèrement plus important que les autres méthodes (3-4%) mais qui reste néanmoins modéré.

Sur le domaine non échantillonné, les estimateurs de Horvitz-Thompson et par calage ne peuvent pas être produits. L'écart entre la meilleure méthode (Sinha-Rao) et les autres est encore plus marqué. On constate par ailleurs que l'ensemble des méthodes ont un biais non négligeable, de l'ordre de 10%. Ce domaine semble donc différent des autres, même une fois prises en compte les variables explicatives. L'estimateur de Sinha-Rao est celui pour lequel le biais est le plus modéré, ce qui explique ses bonnes performances. Pour ces domaines, les modèles linéaires mixtes arrivent en seconde position, suivis par les méthodes non paramétriques et ensuite les régressions linéaires.

De manière plus générale, l'application d'approches robustes basées sur des mesures d'influence améliore légèrement la qualité des estimations pour les méthodes

non paramétriques ainsi que pour les modèles linéaires mixtes dans le cas des domaines non échantillonnés. De manière surprenante, cette approche n'améliore en revanche pas la qualité de l'estimation par modèles linéaires mixtes pour les domaines échantillonnés.

En outre, on constate que les modèles linéaires mixtes sont globalement meilleurs que les régressions linéaires. On peut donc supposer qu'il existe des spécificités des domaines non explicables par les seules variables auxiliaires, ce qui fait que les modèles linéaires mixtes surpassent légèrement les régressions linéaires. De même, il est probable que la forêt aléatoire permette de capter des non linéarités dans la relation entre variables explicatives et courbes qui ne sont pas traduites par le modèle linéaire mixte, ce qui explique ces bonnes performances.

Que ce soit sur les domaines robustes ou non robustes, le fait de projeter les courbes sur la base des composantes principales ou des ondelettes ne semble pas apporter de gain de précision conséquent ici. De manière générale, la forêt aléatoire donne des résultats légèrement meilleurs que les arbres de régression. Par ailleurs, sur ce jeu de données particulier, la méthode Courbotree donne de meilleurs résultats lorsque l'on travaille sur les courbes brutes et non pas lorsque l'on distingue l'estimation de la forme et du niveau (en effet, la normalisation semble globalement augmenter le biais relatif).

Enfin, au niveau des temps de calcul, comme attendu, la régression linéaire fonctionnelle implémentée en utilisant l'algorithme du calage est beaucoup plus rapide que les modèles linéaires mixtes car elle ne nécessite de calculer qu'un unique jeu de poids quelle que soit la longueur des courbes considérées. Cependant, cette méthode est moins performante que les autres. Enfin, les forêts aléatoires sont évidemment beaucoup plus gourmandes en temps de calcul que les arbres de régression. Les versions robustes des méthodes sont globalement beaucoup plus lentes. Ainsi, l'arbre de régression robuste est extrêmement lent (3 minutes) puisqu'il faut réestimer un arbre par individu de l'échantillon. Pour la version robuste de la forêt aléatoire, le temps de calcul est également très important (2 minutes). Probablement du fait d'une implémentation très efficace dans le package R rsae, l'estimateur de Sinha Rao est très rapide, même plus que la version non robuste du modèle linéaire mixte (2 secondes contre 5 en version non robuste).

type domaine	méthode	RE (%)	MSE (10e3)	RB (%)
échantillonné	naïf	100.00	46.27	0.42
échantillonné	calage	126.91	59.25	0.69
échantillonné	calage robuste	125.85	59.26	-1.06
échantillonné	courboforest	26.48	12.75	3.24
échantillonné	courboforest robuste	**25.80**	12.37	3.09
échantillonné	courbotree	30.21	14.64	3.51
échantillonné	courbotree normalisé	33.51	15.78	4.27
échantillonné	courbotree robuste	29.49	14.29	3.01
échantillonné	modèle linéaire mixte	28.61	13.57	1.84
échantillonné	modèle linéaire mixte ondelettes	29.72	13.61	1.72
échantillonné	modèle linéaire mixte robuste	31.58	15.01	0.99
échantillonné	modèle linéaire mixte robuste ondelettes	33.36	15.07	0.98
échantillonné	modèle linéaire mixte robuste sur acp	31.44	14.93	0.97
échantillonné	modèle linéaire mixte sur acp	28.38	13.48	1.75
échantillonné	régression linéaire	30.47	14.67	3.66
échantillonné	régression linéaire robuste	30.44	14.66	3.58
échantillonné	régression linéaire robuste acp	30.44	14.66	3.59
échantillonné	sinha rao	**25.55**	11.96	-1.34
échantillonné	sinha rao sur acp	**24.21**	11.46	-0.10
non échantillonné	naïf			
non échantillonné	calage			
non échantillonné	calage robuste			
non échantillonné	courboforest	40.63	19.04	11.96
non échantillonné	courboforest robuste	39.48	18.44	11.91
non échantillonné	courbotree	44.13	21.07	11.91
non échantillonné	courbotree normalisé	59.14	26.64	14.21
non échantillonné	courbotree robuste	39.94	18.97	11.46
non échantillonné	modèle linéaire mixte	34.52	15.77	10.24
non échantillonné	modèle linéaire mixte ondelettes	33.96	15.66	10.10
non échantillonné	modèle linéaire mixte robuste	**29.97**	13.62	9.02
non échantillonné	modèle linéaire mixte robuste ondelettes	**30.14**	13.66	9.04
non échantillonné	modèle linéaire mixte robuste sur acp	29.67	13.57	8.99
non échantillonné	modèle linéaire mixte sur acp	33.97	15.59	10.14
non échantillonné	régression linéaire	42.26	19.57	12.00
non échantillonné	régression linéaire robuste	41.84	19.37	11.92
non échantillonné	régression linéaire robuste acp	41.87	19.38	11.93
non échantillonné	sinha rao	**21.96**	10.41	6.66
non échantillonné	sinha rao sur acp	**25.68**	11.39	7.95

TABLEAU 4.3 – Moyennes des indicateurs de performances des méthodes (RB, MSE, RE), pour l'ensemble des instants de discrétisation et des domaines, en séparant le domaine non échantillonné des autres.

methode	temps de calcul (secondes)
naïf	0.08
calage	0.21
calage robuste	2.96
courboforest	31.02
courboforest robuste	119.73
courbotree	0.79
courbotree normalisé	0.79
courbotree robuste	169.69
modèle linéaire mixte	5.68
modèle linéaire mixte ondelettes	8.40
modèle linéaire mixte robuste	17.08
modèle linéaire mixte robuste ondelettes	22.82
modèle linéaire mixte robuste sur acp	17.58
modèle linéaire mixte sur acp	6.04
régression linéaire	0.07
régression linéaire robuste	1.20
régression linéaire robuste acp	1.47
sinha rao	2.09
sinha rao sur acp	2.28

TABLEAU 4.4 – Temps de calcul moyen en secondes des différents estimateurs pour un échantillon de 200 courbes de 177 points

4.6 Conclusions sur les petits domaines

4.6.1 Conclusions méthodologiques

Dans ce chapitre, nous avons proposé quatre estimateurs non robustes de courbes moyennes par sondage pour des petites sous-populations. L'ensemble de ces estimateurs relèvent de l'approche basée sur un modèle en sondages. Les deux premières méthodes consistent, d'une façon très similaire à ce qui a été fait au Chapitre précédent, à projeter les courbes dans un espace de dimension finie puis à employer des méthodes usuelles d'estimation de totaux de variables réelles sur chacun des vecteurs de base de l'espace de projection. En l'occurrence, on utilise soit les modèles linéaires mixtes au niveau unité, soit la régression linéaire fonctionnelle, dont les coefficients sont estimés de manière rapide à l'aide d'un estimateur par calage en suivant la démarche proposée par Ardilly (2014). La troisième et la quatrième méthodes consistent à prédire chacune des courbes des unités non échantillonnées par un modèle non paramétrique puis à agréger ces prédictions pour en déduire les courbes moyennes estimées pour chaque domaine. Les modèles utilisés sont des arbres de régression adaptés à des données fonctionnelles construits suivant l'approche Courbotree de Stéphan and Cogordan (2009) ou des forêts aléatoires adaptées à des données fonctionnelles obtenues en agrégeant des arbres Courbotree aléatoires.

Nos tests ont montré que, parmi les méthodes non robustes, les forêts aléatoires donnaient globalement les meilleurs résultats, et permettaient de diviser l'erreur commise par quatre environ. Les modèles linéaires mixtes arrivent ensuite, suivies par les arbres de régression et enfin les régressions linéaires fonctionnelles. Le fait de projeter les courbes sur des bases d'ondelettes ou de composantes principales ne semble alors pas induire de gain de précision notable.

Tout comme dans le cadre du Chapitre précédent, il est possible que nos échantillons contiennent des unités influentes, qui peuvent rendre instables nos estimateurs. Nous avons donc proposé des versions robustes de chacun de nos estimateurs, construites en appliquant la démarche basée sur le biais conditionnel pour des données fonctionnelles décrite dans le Chapitre précédent ou encore en appliquant l'estimateur de Sinha-Rao sur les courbes discrétisées. Cette dernière approche est celle qui fournit les meilleurs résultats, pour des temps de calcul raisonnables. Cependant, dans des cas où les erreurs ne sont pas symétriques, elle peut conduire à des biais conséquents et doit donc être considérée avec prudence. Par ailleurs, l'approche robuste basée sur le biais conditionnel améliore légèrement la performances des estimateurs non paramétriques mais au prix de temps de calculs sensiblement plus importants. Cette amélioration est moins nette pour les modèles linéaires mixtes.

L'ensemble des estimateurs de courbes moyennes par domaines présentés dans ce chapitre ont été implémentés dans un package R interne à EDF, dans leurs versions robustes et non robustes. Il est possible de réaliser des projections sur des bases d'ondelettes, ou sur la base de l'ACP (sphérique pour les méthodes robustes).

4.6.2 Perspectives

Ces travaux peuvent être prolongés de différentes manières. En particulier, l'approche basée sur l'agrégation d'estimations non paramétriques de courbes indivi-

duelles par arbres de régression ou forêts aléatoires nous semble prometteuse. Une piste d'amélioration intéressante pourrait être d'utiliser des distances plus pertinentes que la distance euclidienne dans le critère de split qui permet de construire nos arbres de régression. On pourrait ainsi utiliser la distance de Mahalanobis, la distance de Manhattan, ou encore une distance de type "dynamic time warping".

Une autre piste pourrait être de construire ce critère de split en appliquant la distance euclidienne non pas sur les courbes discrétisées mais sur une transformation de ces courbes, par projection dans une base d'ondelettes, ou encore sur des résumés non linéaires tels que des autoencodeurs variationnels issus de modèles de deep learning (voir par exemple LeCun et al. (2015)).

En outre, on peut également se poser la question du choix de la profondeur de l'arbre de régression, de la taille minimale des feuilles ainsi que du nombre d'arbres de la forêt. En effet, les critères usuellement utilisés en statistique non paramétrique pour répondre à cette question se basent généralement sur le principe de la validation croisée. Cependant, notre objectif n'est pas ici de déterminer la meilleure prédiction possible pour chaque unité de la population, mais bien la prédiction qui aboutit à la meilleure estimation de courbe moyenne par domaine, ce qui n'est pas obligatoirement la même chose. Il serait donc souhaitable d'adapter les critères de validation croisée de façon à refléter notre objectif.

Il serait également intéressant de réaliser nos prédictions individuelles par d'autres méthodes non paramétriques basées sur des arbres de régression, par exemple les Gradient Boosted Machine (Ye et al. (2009)) ou les arbres XGboost (Chen and Guestrin (2016)) que l'on adapterait à la prévision de données fonctionnelles en suivant une approche similaire à celle du Courbotree (utilisation d'une distance fonctionnelle dans le critère de split pour la construction des arbres).

Enfin, on remarque que l'introduction d'effets aléatoires dans les modèles linéaires induit une amélioration de la prédiction, ce qui nous amène à penser qu'il existe des spécificités des domaines qui ne s'expliquent pas uniquement par les informations auxiliaires. Il pourrait donc être pertinent d'adapter les arbres de régression fonctionnels de façon à inclure des effets aléatoires. Un solution serait d'étendre par exemple l'algorithme de Hajjem et al. (2014), basé sur un algorithme EM, au cadre des données fonctionnelles.

Chapitre 5

Estimation de courbes moyennes par sondage en présence de valeurs manquantes

Dans ce chapitre, on s'intéresse à l'estimation de courbes totales ou moyennes par sondage en présence de courbes partiellement observées. En effet, du fait de problèmes techniques pouvant survenir tout au long de la chaîne d'acquisition, de collecte et de stockage de la donnée, il est possible que certaines portions de certaines courbes de nos panels soient manquantes.

5.1 Introduction et contexte

Dans ce chapitre, on se place dans le contexte déjà évoqué dans le chapitre 3 : on souhaite estimer la courbe moyenne μ_Y ou totale t_Y d'une population d'intérêt U à partir d'un échantillon de courbes sélectionné selon un plan de sondage aléatoire. Dans le cadre de EDF, cette population peut être un ensemble de clients partageant des caractéristiques communes (équipements électriques, type de contrat, lieu de résidence par exemple).

Jusqu'à présent, nous avions supposé que l'intégralité des courbes de nos panels étaient connues pour l'ensemble des instants de discrétisation de la période d'étude. Malheureusement en pratique, du fait de difficultés techniques pouvant survenir tout au long de la chaîne de mesure, de remontée et de stockage des données, il est possible que certaines portions de courbes soient manquantes. Ce problème s'apparente à celui de la non réponse dans le cadre usuel des sondages et la qualité des estimations de moyennes ou de totaux peut être dégradée. Dans la suite, par abus de langage, on parlera donc de non réponse pour désigner les valeurs manquantes. Il peut s'agir de non réponse partielle lorsque seules certaines portions de courbes sont manquantes ou totale lorsque l'ensemble d'une courbe est manquante. La structure de nos valeurs manquantes est complexe, avec des séquences manquantes de type "gruyère" c'est-à-dire des séquences de longueurs variées, survenant et se terminant à des instants différents pour les différents individus.

Bien qu'il existe une importante littérature en sondages sur l'estimation en présence de valeurs manquantes (voir par exemple, Särndal and Lundström (2005), ou Ha-

ziza (2009) ainsi que les références citées dans ces articles), les techniques usuelles sont en général univariées et ne permettent donc pas de prendre en compte l'aspect fonctionnel d'un problème tel que le nôtre (ou en tout cas les corrélations fortes entre les valeurs des courbes aux différents instants). En présence de valeurs manquantes, on peut chercher à atteindre deux objectifs, qui requièrent chacun l'emploi de méthodes spécifiques. Le premier est de reconstituer du mieux possible chacune des portions de courbes manquantes : plutôt qu'un problème de sondage il s'agit alors d'un problème classique de modélisation statistique, pour des variables cibles fonctionnelles et des variables explicatives scalaires ou fonctionnelles (on pourra alors se référer aux ouvrages de référence sur les données fonctionnelles précédemment cités : Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012) et Cuevas (2014). Ce premier problème ne sera pas celui qui nous motivera ici. Dans ce chapitre, nous cherchons plutôt à estimer du mieux possible la courbe moyenne (ou totale) pour l'ensemble de la population à partir de notre échantillon de courbes dont certaines pourront être partiellement ou totalement inobservées.

Quatre approches sont proposées et comparées. La première, présentée dans 5.2, se base sur un lissage temporel non paramétrique inspiré de Staniswalis and Lee (1998) et adapté aux estimations par sondage en présence de valeurs manquantes dans Cardot et al. (2015). Dans cette approche, on approxime la courbe moyenne par son estimateur par repondération par lissage. Une approximation de la variance de cet estimateur est proposée dans 5.2.2. On introduit ensuite trois méthodes d'imputation, c'est-à-dire des techniques consistant à remplacer les valeurs manquantes par des valeurs jugées pertinentes. La première de ces méthodes, l'imputation par l'estimateur à noyau, est présentée dans 5.3.1 et consiste à remplacer les valeurs manquantes par les valeurs de l'estimateur par repondération précédemment évoqué. Une deuxième approche, présentée dans 5.3.2, consiste à réaliser, pour chaque portion de courbe manquante, une imputation par donneur basée sur les plus proches voisins (voir Chen and Shao (2000) ou Beaumont and Bocci (2009)). Cela revient à supposer implicitement l'existence d'un modèle non paramétrique traduisant la relation entre la portion inobservée des trajectoires et les variables explicatives (comportant entre autres les portions observées de ces mêmes trajectoires). En raison du compromis entre biais et variance, usuel en statistique non paramétrique, on cherche à garder un nombre de variables explicatives le plus petit possible et le choix de celles-ci sera donc un enjeu important. Pour ces deux premières approches on suit une approche mixte qui tient compte d'un modèle de non réponse ainsi que d'un modèle d'imputation (voir par exemple Beaumont and Bissonnette (2011)).

En pratique, pour ces méthodes, il est de coutume de constituer au préalable des classes d'imputation à l'aide des informations auxiliaires, puis de réaliser l'imputation au sein de chaque classe. Si l'information auxiliaire est pertinente, on s'attend en effet à ce que les valeurs de la variable d'intérêt soient plus homogènes au sein d'une classe d'imputation, ce qui apporte une certaine stabilité à la méthode (voir par exemple Haziza (2005)). Dans notre contexte d'imputation de courbes, comme nous le verrons dans 5.3.2, cela permet en outre de combiner les informations de "long terme" sur les individus, utilisées pour constituer ces classes, et les informations sur les instants entourant chacune des séquences de valeurs manquantes qui sont utilisées pour choisir les voisins au sein de chaque classe. En outre, l'usage de classes d'imputation, dont

on peut augmenter le nombre à mesure que la taille de l'échantillon grandit, peut permettre de limiter les temps de calcul pour de gros volumes de données.

Enfin, une dernière technique, propre aux données fonctionnelles suffisamment régulières, consiste à réaliser pour chaque séquence de valeurs manquantes une interpolation linéaire entre le premier et le dernier instant observés. Intuitivement, on se doute que cette méthode, très simple et très employée en pratique, peut se révéler efficace sur les séquences de valeurs manquantes courtes. On considère dans 5.3.3 une variante de cette méthode un peu plus évoluée qui consiste à interpoler l'écart de la trajectoire à la courbe moyenne puis à l'additionner à cette courbe moyenne. Cette variante permet de capter l'évolution globale de la courbe de la population sur la période considérée. La variance de ces estimateurs par imputation est approximée en suivant une démarche unifiée que nous détaillons dans 5.3.4.

Enfin, dans la section 5.4, ces différents estimateurs ainsi que leurs estimateurs de variance sont testés et comparés entre eux sur le jeu de données des courbes irlandaises déjà présenté dans le chapitre 3.

La partie de ce chapitre concernant l'estimation de courbe moyenne par lissage (5.2) est tirée de l'article Cardot et al. (2015) et les parties 5.3 et 5.4 concernant les méthodes par imputation ainsi que les tests sur des données réelles ont quant à elles fait l'objet d'un article soumis pour un numéro spécial de la Revue Canadienne de Statistique.

5.1.1 Notations

On se place dans le cadre de travail décrit dans la section 2.2.1. Plus précisément, on souhaite estimer la courbe moyenne de la population :

$$\mu_Y(t) = \frac{1}{N}\sum_{i\in U} Y_i(t), \quad t \in [0,T], \tag{5.1}$$

ou encore la courbe totale

$$t_Y(t) = N\mu_Y(t), \quad t \in [0,T]. \tag{5.2}$$

On se place dans l'approche basée sur le plan. Les indicatrices d'appartenance à l'échantillon $\mathbb{1}_i = \mathbb{1}_{i\in s}$, $i \in U$ sont donc des variables aléatoires et les valeurs de la variable d'intérêt Y_i sont considérées comme déterministes.

Lorsqu'il n'y a pas de non réponse, la courbe totale t_Y de la population U peut être estimée par l'estimateur de Horvitz-Thompson

$$\hat{t}_Y(t) = \sum_{i\in s} d_i Y_i(t), \quad t \in [0,T], \tag{5.3}$$

avec $d_i = \frac{1}{\pi_i}$. Du fait de la non réponse, certaines courbes peuvent être partiellement ou totalement inobservées. Pour chaque unité i de l'échantillon, on définit un processus de non réponse $r_i(t)$ dont la valeur est 0 si la courbe Y_i n'est pas observée à l'instant t et 1 sinon. La distribution des r_i peut être variable en fonction du temps et de l'unité considérée.

Pour chaque instant $t \in \{t_1, \ldots, t_L\}$, on note $s_m(t) \subset s$ l'ensemble des non-répondants *i.e* $s_m(t) = \{i \in s \mid r_i(t) = 0\}$ et $s_r(t)$ l'ensemble des répondants $s_r(t) = s - s_m(t)$.

Dans notre contexte, la non-réponse partielle résulte de problèmes techniques et non d'une volonté de l'unité échantillonnée, on peut donc aisément faire l'hypothèse de mécanisme de réponse ignorable [1] (voir Rubin (1976)). Cette hypothèse signifie que, conditionnellement à l'information auxiliaire $\mathbf{X}_i$, la distribution de la variable d'intérêt dans l'échantillon est identique à celle qui prévaudrait dans l'échantillon en l'absence de valeurs manquantes. Pour nos données fonctionnelles, on a donc

$$\mathrm{P}r(r_i(t) = 1|\mathrm{Y}_i, \mathbf{X}_i) = \mathrm{P}r(r_i(t) = 1|\mathbf{X}_i) \quad \forall t \in [0, \mathrm{T}].$$

Afin d'améliorer la performance des différents estimateurs et de limiter les temps de calcul, la population U est partitionnée en C classes d'imputation disjointes (obtenues en fonction de l'information auxiliaire mais aussi en tenant compte de la longueur de la séquence manquante éventuelle) de telle manière que dans chaque classe d'imputation la forme et le niveau des trajectoires soit aussi similaires que possible. On note $\mathrm{U}_c, c = 1, \ldots, \mathrm{C}$ les classes d'imputation, qui satisfont $\mathrm{U} = \bigcup_{c=1}^{\mathrm{C}} \mathrm{U}_c$ et $\mathrm{U}_c \cap \mathrm{U}_{c'} = \emptyset$ pour $c \neq c'$. Ces classes d'imputation peuvent varier au cours du temps, notamment pour les trajectoires comportant plusieurs séquences de valeurs manquantes. [2] On notera en outre N_c le nombre d'unités de la population appartenant à la classe d'imputation U_c, pour $c = 1, \ldots, \mathrm{C}$ et $s_c = s \cap \mathrm{U}_c$ l'ensemble des unités de l'échantillon qui appartiennent à la classe d'imputation U_c.

5.2 Estimation de courbe moyenne ou totale par repondération par lissage à noyau

5.2.1 Estimateur de courbe moyenne par repondération par lissage à noyau

La première stratégie pour estimer la courbe moyenne ou totale consiste à utiliser un estimateur à noyau. Dans cette section, l'estimateur à noyau constituera notre estimateur de courbe moyenne tandis que dans la sous-section 5.3.1 il sera utilisé dans chaque classe d'imputation pour imputer pour les portions de courbes manquantes. Pour proposer cet estimateur, nous allons partir de la méthode d'estimation non paramétrique pour des données fonctionnelles proposée par Staniswalis and Lee (1998) et l'étendre au contexte des estimations par sondage en population finie en présence de trajectoires partiellement observées.

1. On préfère l'hypothèse Missing at Random plutôt que Completely Missing at Random : en effet, la probabilité de non réponse peut dépendre des caractéristiques de l'unité, par exemple du type de compteur communicant installé, qui peut potentiellement être indirectement lié au niveau de consommation, néanmoins on suppose que cette information est connue.

2. En effet, la classe d'imputation dépend de la longueur des séquences manquantes éventuelles, une même courbe peut donc changer de classe d'imputation si elle présente des séquences manquantes de tailles différentes.

Lissage à noyau de la courbe moyenne sur l'ensemble de la population

Lorsqu'il n'y a pas de non réponse, on peut approximer la courbe moyenne μ_Y pour chaque instant $t \in [0, T]$ en appliquant un lisseur à noyau (voir Staniswalis and Lee (1998)), que l'on note $\tilde{\mu}_Y(t)$ ou simplement $\tilde{\mu}(t)$.

Pour cela, on introduit un noyau K(.), *i.e* une fonction continue, positive et symétrique en zéro (voir par exemple Hart (2013) pour une définition plus précise ainsi que des exemples). Des exemples classiques de noyaux sont le noyau gaussien défini par $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ et le noyau d'Epanechnikov décrit notamment dans Wand and Jones (1995)) et défini par $K(x) = \frac{3}{4}(1-x^2)\mathbb{1}_{\{|x|\leq 1\}}$. Pour tout instant $t \in [0, T]$, l'utilisation du noyau nous donne l'approximation lisse suivante de $\mu_Y(t)$,

$$\tilde{\mu}(t) = \frac{\sum_{i=1}^{N}\sum_{l=1}^{L} K\left(\frac{t-t_l}{\eta}\right) Y_i(t_l)}{\sum_{i=1}^{N}\sum_{l=1}^{L} K\left(\frac{t-t_l}{\eta}\right)}, \tag{5.4}$$

avec η une fenêtre permettant de contrôler le caractère lisse de l'approximation. A mesure que η grandit, l'estimateur $\tilde{\mu}$ devient de plus en plus lisse, avec plus de biais mais moins de variance alors que des valeurs plus faibles de η donnent des estimateurs qui peuvent osciller davantage, avec un faible biais mais une plus grande variabilité. Comme les instants de discrétisation sont les mêmes pour l'ensemble des individus, l'expression devient

$$\tilde{\mu}(t) = \sum_{l=1}^{L} w(t, t_l, \eta)\mu(t_l), \tag{5.5}$$

avec les poids de lissage

$$w(t, t_l, \eta) = \frac{K\left(\frac{t-t_l}{\eta}\right)}{\sum_{l=1}^{L} K\left(\frac{t-t_l}{\eta}\right)},$$

donc $\tilde{\mu}(t)$ est simplement obtenu en lissant la trajectoire moyenne de la population $(\mu(t_1), \ldots, \mu(t_L))$ aux instants $(t_1, \ldots, t_L)$.

Estimateurs non paramétriques en sondages

En l'absence de non réponse, un premier estimateur simple de Horvitz-Thompson de l'approximation lisse $\tilde{\mu}(t)$ est obtenu en remplaçant les valeurs moyennes inconnues $\mu(t_l)$ dans (5.5) par leurs estimateurs de Horvitz-Thompson $\widehat{\mu}(t_l)$. On obtient alors, pour $t \in [0, T]$,

$$\widehat{\mu}_{HT}(t) = \sum_{l=1}^{L} w(t, t_l, \eta)\, \widehat{\mu}(t_l), \tag{5.6}$$

où

$$\widehat{\mu}(t_l) = \frac{1}{N}\sum_{i\in s} \frac{Y_i(t_l)}{\pi_i}. \tag{5.7}$$

Une idée alternative est d'estimer $\tilde{\mu}_Y(t)$ par un estimateur de type Hájek, obtenu en remplaçant les $\mu(t_l)$ par leurs estimateurs de Hájek.

$$\widehat{\mu}_{Y,Ha}(t) = \frac{\sum_{i\in s}\sum_{l=1}^{L} K\left(\frac{t-t_l}{\eta}\right)\frac{Y_i(t_l)}{\pi_i}}{\sum_{i\in s}\sum_{l=1}^{L} K\left(\frac{t-t_l}{\eta}\right)\frac{1}{\pi_i}} \tag{5.8}$$

$$= \sum_{l=1}^{L} w(t,t_l,\eta)\left(\frac{\sum_{i\in s}\frac{Y_i(t_l)}{\pi_i}}{\sum_{i\in s}\frac{1}{\pi_i}}\right). \tag{5.9}$$

L'intérêt de cette nouvelle écriture est alors de prendre en compte la taille de la population dans l'estimation de la moyenne. Quelle que soit l'estimateur considéré, on en déduit ensuite naturellement l'estimateur de la courbe totale lissée $\widehat{\tilde{t}}_Y(t)$ en le multipliant par la taille de la population N.

Estimateurs non paramétriques en sondage en présence de non réponse

Comme évoqué précédemment, certaines portions de certaines courbes de l'échantillon peuvent être manquantes. Pour chaque unité i de l'échantillon, on note

$$\vartheta_i(t_l) = Pr\left(r_i(t_l) = 1\right),$$

la probabilité que la courbe de l'unité i soit observée à l'instant t_l et

$$\vartheta_i(t_l, t_{l'}) = Pr\left(r_i(t_l) = 1 \;\&\; r_i(t_{l'}) = 1\right),$$

la probabilité que la courbe de l'unité i soit observée aux instants t_l et $t_{l'}$. Pour simplifier les calculs, on supposera qu'il existe un petit nombre de groupes homogènes de réponse (voir Lundström and Särndal (1999)) c'est-à-dire des groupes au sein desquels les schémas de non réponse des unités sont identiques (avec les mêmes probabilités de non réponse pour chaque instant ou couple d'instants) et indépendants. Par exemple, dans notre contexte, ces groupes de non réponse peuvent correspondre aux caractéristiques techniques des compteurs installés chez les individus du panel.

On peut ensuite définir trois estimateurs différents de $\tilde{\mu}(t)$ basés sur le lissage et la repondération et qui prennent en compte la non réponse. Le premier, dérivé de (5.6), est un estimateur de Horvitz-Thompson lissé qui intègre la non réponse,

$$\widehat{\mu}_{r,HT}(t) = \frac{1}{N}\sum_{l=1}^{L} w(t,t_l,\eta)\left(\sum_{i\in s}\frac{r_i(t_l)Y_i(t_l)}{\vartheta_i(t_l)\pi_i}\right). \tag{5.10}$$

Le second est dérivé de (5.8), et peut être vu comme un estimateur de type Hájek de deux ratios lissés,

$$\widehat{\mu}^{(1)}_{r,Ha}(t) = \frac{\sum_{l=1}^{L} K(\eta^{-1}(t-t_l))\left(\sum_{i\in s}\frac{r_i(t_l)}{\vartheta_i(t_l)}\frac{Y_i(t_l)}{\pi_i}\right)}{\sum_{l=1}^{L} K(\eta^{-1}(t-t_l))\left(\sum_{i\in s}\frac{r_i(t_l)}{\vartheta_i(t_l)}\frac{1}{\pi_i}\right)}. \tag{5.11}$$

Le dernier estimateur est également un estimateur de Hájek lissé et provient de (5.9). On le définit, pour $t \in [0, \mathrm{T}]$, par

$$\widehat{\mu}^{(2)}_{r,\mathrm{H}a}(t) = \sum_{l=1}^{\mathrm{L}} w(t, t_l, \eta) \frac{\widehat{\mathrm{Y}}(t_l)}{\widehat{\mathrm{N}}(t_l)}, \tag{5.12}$$

avec $\widehat{\mathrm{Y}}(t_l) = \sum_{i \in s} \frac{r_i(t_l)}{\vartheta_i(t_l)} \frac{\mathrm{Y}_i(t_l)}{\pi_i}$ et $\widehat{\mathrm{N}}(t_l) = \sum_{i \in s} \frac{r_i(t_l)}{\vartheta_i(t_l)} \frac{1}{\pi_i}$. En l'absence de non réponse, les estimateurs $\widehat{\mu}^{(1)}_{r,\mathrm{H}a}(t)$ et $\widehat{\mu}^{(2)}_{r,\mathrm{H}a}(t)$ sont identiques. Dans la suite, on s'intéresse plus particulièrement à la seconde version de l'estimateur de Hájek (5.11). Cet estimateur sera celui utilisé pour imputer les valeurs manquantes dans la sous-section 5.3.1.

Estimation des probabilités de réponse $\vartheta_c(t_l)$ et $\vartheta_c(t_l, t_{l'})$.

Dans le cas général pour lequel chaque unité i est affectée par un mécanisme de non réponse différent et potentiellement dépendant du temps, il devient presque impossible d'estimer la probabilité de réponse. On suppose donc que la population U est découpée en groupes de réponses homogènes (qui seront aussi les classes d'imputation par la suite). On peut alors considérer les estimateurs de Horvitz-Thompson de ces probabilités de réponse. Dans la sous-population U_c,

$$\widehat{\vartheta}_c(t_l) = \frac{1}{\mathrm{N}} \sum_{i \in s_c} \frac{1}{\pi_i} r_i(t_l)$$

et

$$\widehat{\vartheta}_c(t_l, t_{l'}) = \frac{1}{\mathrm{N}} \sum_{i \in s_c} \frac{1}{\pi_i} r_i(t_l) r_i(t_{l'}).$$

On peut supposer également que le processus de non réponse est stationnaire au second ordre, c'est-à-dire que $\widehat{\vartheta}_c(t_l)$ ne dépend pas de t_l et $\widehat{\vartheta}_c(t_l, t_{l'})$ ne dépend que de $|t_l - t_{l'}|$, on en déduit donc les estimateurs suivants

$$\widehat{\vartheta}_c = \frac{1}{\mathrm{N}} \sum_{i \in s_c} \frac{1}{\pi_i} \sum_{l=1}^{\mathrm{L}} r_i(t_l)$$

et, pour chaque couple $(t_l, t_{l'})$ tel que $|t_l - t_{l'}| = \Delta_t$,

$$\widehat{\vartheta}_c(\Delta_t) = \frac{1}{\mathrm{N}} \sum_{i \in s_i} \frac{1}{\pi_i} \sum_{l, l' | t_l - t_{l'} | = \Delta_t} r_i(t_l) r_i(t_{l'}).$$

Ces estimateurs peuvent être implémentés soit directement sur le jeu de données utilisé pour l'estimation de la courbe moyenne soit sur un autre jeu de données plus gros à condition que celui-ci ait été collecté par des compteurs possédant les mêmes caractéristiques techniques.

Cas particulier du sondage stratifié avec groupes de réponses homogènes par strate

Dans cette sous-section, on considère le cas particulier fréquent du sondage stratifié, avec sondage aléatoire simple dans chaque strate. On va considérer de plus que ces strates constituent des groupes de réponse homogènes. La population U est donc

divisée en H strates, $U_h, h = 1, \ldots, H$, de tailles N_H telles que $U = \bigcup_{h=1}^{H} U_h$, $U_h \bigcap U_{h'} = \emptyset$ pour $h \neq h'$ et $N = \sum_{h=1}^{H} N_h$. Les H strates sont construites en fonction d'une information auxiliaire liée à la forme et au niveau des courbes. La courbe moyenne s'écrit alors

$$\mu_Y = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h \tag{5.13}$$

où, pour chaque $h \in \{1, \ldots, H\}$, μ_h est la courbe moyenne de la strate U_h,

$$\mu_h = \frac{1}{N_h} \sum_{i \in U_h} Y_i. \tag{5.14}$$

Différents lissages à noyaux peuvent être employés dans chaque strate et l'approximation globale de μ par lissage à noyau est finalement obtenue par combinaison linéaire,

$$\widetilde{\mu} = \sum_{h=1}^{H} \frac{N_h}{N} \widetilde{\mu}_h \tag{5.15}$$

où, pour chaque h et chaque $t \in [0, T]$,

$$\widetilde{\mu}_h(t) = \sum_{l=1}^{L} w(t, t_l, \eta) \mu_h(t_l), \tag{5.16}$$

avec les poids de lissage $w(t, t_l, \eta)$ définis dans (5.5) qui peuvent être différents d'une strate à l'autre.

Dans chaque strate h, le mécanisme de non réponse est supposé homogène et la probabilité de réponse est la même pour chaque unité au sein d'une strate : pour $i \in U_h$, alors $Pr(r_i(t_l) = 1) = \vartheta_h(t_l)$.

Pour chaque strate h, on peut écrire, comme précédemment, trois estimateurs de $\widetilde{\mu}_h$. En utilisant les expressions (5.11) et (5.12), on obtient pour $t \in [0, T]$ et $s_h = s \cap U_h$,

$$\widehat{\mu}^{(1)}_{h,\eta,Ha}(t) = \frac{\sum_{l=1}^{L} K(\eta^{-1}(t - t_l)) \left(\sum_{i \in s_h} \frac{Y_i(t_l)}{\pi_i} \frac{r_i(t_l)}{\vartheta_h(t_l)} \right)}{\sum_{l=1}^{L} K(\eta^{-1}(t - t_l)) \left(\sum_{i \in s_h} \frac{1}{\pi_i} \frac{r_i(t_l)}{\vartheta_h(t_i)} \right)}, \tag{5.17}$$

et

$$\widehat{\mu}^{(2)}_{h,\eta,Ha}(t) = \sum_{l=1}^{L} w(t, t_l, \eta) \left(\frac{\sum_{i \in s_h} \frac{Y_i(t_l)}{\pi_i} \frac{r_i(t_l)}{\vartheta_h(t_l)}}{\sum_{i \in s_h} \frac{1}{\pi_i} \frac{r_i(t_l)}{\vartheta_h(t_l)}} \right), \tag{5.18}$$

ou

$$\widehat{\mu}_{h,\eta,HT}(t) = \sum_{l=1}^{L} w(t, t_l, \eta) \left(\sum_{k \in s_c} \frac{Y_i(t_l)}{\pi_i} \frac{r_i(t_l)}{\vartheta_h(t_l)} \right) \tag{5.19}$$

obtenu à partir de l'estimateur défini par (5.10).

Comme précédemment, on s'intéressera plus particulièrement au second estimateur. De plus, en combinant les estimateurs par strates, on en déduit un estimateur de la courbe totale de la population entière par lissage en présence de non réponse :

$$\widehat{t}_Y(t) = \sum_{h=1}^{C} N_h \widehat{\mu}^{(2)}_{h,\eta,Ha}(t), \quad t \in [0, T].$$

Choix de la fenêtre

Comme souvent en statistique non paramétrique, le choix de la fenêtre η de l'estimateur à noyau est capital pour obtenir un estimateur du total aussi précis que possible. Une erreur classique dans ce contexte non paramétrique serait de choisir cette fenêtre en utilisant un critère de validation croisée standard, conçu pour estimer du mieux possible chacune des trajectoires et non pas la moyenne des trajectoires. Par exemple, Cardot et al. (2013a) montrent dans un contexte similaire que l'interpolation linéaire peut surpasser le lissage à noyau, même lorsque le niveau de bruit est plutôt élevé, si la valeur de la fenêtre est choisie par un critère de validation croisée usuel appliqué courbe par courbe. Cette procédure de lissage individuel conduit en effet à un lissage global excessif (voir aussi Hart and Wehrly (1993)) et alors le biais de l'estimateur de la moyenne ainsi obtenu est beaucoup plus important que sa variance. Comme dans Cardot et al. (2013a), nous proposons donc ici d'employer un critère de validation croisée reposant sur le principe du leave-one-out, avec des poids modifiés. Ce critère modifié prend en compte à la fois le plan de sondage et le processus de non réponse. La valeur de la fenêtre est choisie de façon à minimiser

$$\mathrm{CV}(\eta) = \sum_{i\in s} \frac{1}{\pi_i} \sum_{l=1}^{\mathrm{L}} \frac{r_i(t_l)}{\vartheta_i(t_l)} \left(\mathrm{Y}_i(t_l) - \widehat{\mu}^{(-i)}(t_l)\right)^2, \tag{5.20}$$

avec $\widehat{\mu}^{(-i)}$ l'estimateur de la trajectoire moyenne $\widetilde{\mu}$ construit à partir de l'échantillon s sans utiliser la courbe Y_i.

En particulier, dans le cas du sondage stratifié, le fait de considérer des paramètres de lissage différents dans chaque strate peut ne pas être plus efficace car cela peut conduire à un lissage excessif. En effet, les meilleures approximations de la courbe moyenne de chaque strate U_h en termes d'erreur quadratique moyenne peuvent ne pas conduire à la meilleure approximation de la courbe moyenne de l'ensemble de la population une fois combinées.

5.2.2 Estimation de variance pour l'estimateur à noyau par repondération

Nous allons maintenant proposer une approximation de la variance pour notre estimateur par repondération par lissage à noyau. La démarche employée ici sera différente de l'approche unifiée utilisée par la suite pour les méthodes par imputation : en effet, contrairement aux autres estimateurs présentés dans ce chapitre les estimateurs par lissage proposés dans les équations (5.10), (5.11) et (5.12) ne consistent pas à imputer les valeurs manquantes mais à repondérer les estimations et lisser les estimateurs pour pouvoir les implémenter en présence de valeurs manquantes.

Ici, on commence par montrer, sous des hypothèses générales, que le biais est négligeable en comparaison de la variance. Il est donc pertinent de se concentrer sur l'estimation de la variance de nos trois estimateurs par lissage. Dans la suite, on note $\mathbb{E}_p$ l'espérance (respectivement $\mathbb{V}_p$ la variance) par rapport au plan de sondage et $\mathbb{E}_{\mathrm{R}}$ l'espérance (resp. $\mathbb{V}_{\mathrm{R}}$ la variance) sous le mécanisme de non réponse. Lorsqu'il n'y a pas d'indice, l'espérance $\mathbb{E}$ et la variance $\mathbb{V}$ intègrent simultanément le plan de sondage et le mécanisme de non réponse.

L'erreur d'approximation et le biais sont négligeables

On considère tout d'abord l'estimateur de Horvitz-Thompson $\widehat{\mu}_{r,\mathrm{HT}}(t)$ défini dans (5.10) et on a clairement

$$\begin{aligned}\mathbb{E}(\widehat{\mu}_{r,\mathrm{HT}}(t)) &= \sum_{l=1}^{\mathrm{L}} w(t,t_l,h)\mathbb{E}\left[\sum_{i\in s}\frac{r_i(t_l)\mathrm{Y}_i(t_l)}{\vartheta_i(t_l)\pi_i}\right]\\ &= \sum_{l=1}^{\mathrm{L}} w(t,t_l,h)\mu(t_l)\\ &= \tilde{\mu}(t). \end{aligned} \tag{5.21}$$

un estimateur sans biais de $\tilde{\mu}(t)$. L'erreur quadratique moyenne satisfait donc

$$\mathbb{E}\left[\widehat{\mu}_{r,\mathrm{HT}}(t)-\mu(t)\right]^2 = \left|\tilde{\mu}(t)-\mu(t)\right|^2 + \mathbb{V}\left(\widehat{\mu}_{r,\mathrm{HT}}(t)\right). \tag{5.22}$$

De plus, on peut montrer sous des conditions de régularité générales sur la trajectoire moyenne données dans l'Annexe 5.5.2 que, si la taille de la population N tend vers l'infini, que la fenêtre η tend vers zéro et que le nombre de points de discrétisation tend vers l'infini en satisfaisant $2\eta > (d-1)^{-1}$, alors l'erreur d'approximation est bornée, pour une constante Λ_t, de la manière suivante

$$\left|\tilde{\mu}(t)-\mu(t)\right| \leq \Lambda_t \eta^{\beta}. \tag{5.23}$$

En combinant (5.22) et (5.23), cela implique que, à condition que $\sqrt{n}\eta^{\beta} \to 0$ lorsque la taille d'échantillon $n \to \infty$, l'erreur d'approximation $\tilde{\mu}(t)-\mu(t)$ est négligeable en comparaison de l'erreur d'échantillonnage, qui tend vers zéro en probabilité au maximum à la vitesse de $1/\sqrt{n}$. On note que la condition sur la fenêtre η, qui doit être petite, et la taille d'échantillon impliquent également que la grille de discrétisation doit être suffisamment dense pour que $\sqrt{n}\max_l|t_{l+1}-t_l|^{\beta} \to 0$. Dans ce cas, l'erreur quadratique moyenne de l'estimateur de Horvitz-Thompson peut être approximée par sa variance

$$\mathbb{E}\left[\widehat{\mu}_{r,\mathrm{HT}}(t)-\mu(t)\right]^2 \approx \mathbb{V}\left(\widehat{\mu}_{r,\mathrm{HT}}(t)\right).$$

Les estimateurs $\widehat{\mu}^{(1)}_{r,\mathrm{H}a}(t)$ et $\widehat{\mu}^{(2)}_{r,\mathrm{H}a}(t)$ ne sont pas des estimateurs sans biais de $\tilde{\mu}(t)$ mais sont des estimateurs par le ratio dont le numérateur et le dénominateur sont sans biais. Néanmoins, ils sont asymptotiquement sans biais pour $\tilde{\mu}(t)$ et sous les conditions précédentes, la somme du carré de leur biais et du carré de leur erreur d'approximation est négligeable en comparaison de leur variance. Leurs erreurs quadratiques moyennes peuvent donc être approximées par leur variances.

Approximation de variance pour l'estimateur de Horvitz-Thompson

Les calculs présentés dans le paragraphe ci-dessous sont fortement inspirés du chapitre 15 de Särndal (1992). Pour commencer, on décompose la variance de $\widehat{\mu}_{r,\mathrm{HT}}(t)$ en utilisant de l'approche renversée (voir Shao and Steel (1999))

$$\mathbb{V}(\widehat{\mu}_{r,\mathrm{HT}}(t)) = \mathbb{V}_{\mathrm{R}}\mathbb{E}_p\left(\widehat{\mu}_{r,\mathrm{HT}}(t)-\tilde{\mu}(t)|s_r\right) + \mathbb{E}_{\mathrm{R}}\mathbb{V}_p\left(\widehat{\mu}_{r,\mathrm{HT}}(t)-\tilde{\mu}(t)|s_r\right), \tag{5.24}$$

où s_r est l'ensemble des répondants pour chacun des instants de discrétisation $\{t_1, \ldots, t_L\}$. Pour simplifier les équations, on désignera $w(t, t_l, \eta)$ par $w_l(t)$. On a

$$\mathbb{E}_p(\widehat{\mu}_{r,\mathrm{HT}}(t) - \tilde{\mu}(t)|s_r) = \frac{1}{N}\sum_{l=1}^{L} w_l(t)\left[\sum_{i\in U} Y_i(t_l)\left(\frac{r_i(t_l)}{\vartheta_i(t_l)} - 1\right)\right]$$

et, par indépendance entre les mécanismes de non réponse pour les différentes unités,

$$\begin{aligned}\mathbb{V}_R\mathbb{E}_p(\widehat{\mu}_{r,\mathrm{HT}}(t) - \tilde{\mu}(t)|s_r) &= \frac{1}{N^2}\sum_{i\in U}\sum_{l=1}^{L} w_l^2(t)Y_i^2(t_l)\frac{1-\vartheta_i(t_l)}{\vartheta_i(t_l)} \\ &+ \frac{1}{N^2}\sum_{i\in U}\sum_{l=1}^{L}\sum_{l'\neq l} w_l(t)w_{l'}(t)Y_i(t_l)Y_i(t_{l'})\frac{\vartheta_i(t_l,t_{l'}) - \vartheta_i(t_l)\vartheta_i(t_{l'})}{\vartheta_i(t_l)\vartheta_i(t_{l'})}.\end{aligned} \tag{5.25}$$

On définit $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$ pour $i \neq j$ et $\Delta_{ii} = \pi_i(1-\pi_i)$. On a

$$\mathbb{V}_p(\widehat{\mu}_{r,\mathrm{HT}}(t) - \tilde{\mu}(t)|s_r) = \frac{1}{N^2}\sum_{i,j\in U}\frac{\Delta_{ij}}{\pi_i\pi_j}\left(\sum_{l=1}^{L} w_l(t)\frac{r_i(t_l)}{\vartheta_i(t_l)}Y_i(t_l)\right)\left(\sum_{l'=1}^{L} w_{l'}(t)\frac{r_j(t_{l'})}{\vartheta_j(t_{l'})}Y_j(t_{l'})\right)$$

et en prenant l'espérance sous le mécanisme de non réponse, on obtient

$$\begin{aligned}\mathbb{E}_R\mathbb{V}_p(\widehat{\mu}_{r,\mathrm{HT}}(t) - \tilde{\mu}(t)|s_r) &= \frac{1}{N^2}\sum_{i\in U}\frac{1-\pi_i}{\pi_i}\sum_{l,l'\neq l}^{L} w_l(t)w_{l'}(t)Y_i(t_l)Y_i(t_l')\frac{\vartheta_i(t_l,t_{l'})}{\vartheta_i(t_l)\vartheta_i(t_{l'})} \\ &+ \frac{1}{N^2}\sum_{i\in U}\frac{1-\pi_i}{\pi_i}\sum_{l=1}^{L} w_l^2(t)Y_i^2(t_l)\frac{1}{\vartheta_i(t_l)} \\ &+ \frac{1}{N^2}\sum_{i\in U}\sum_{j\neq i\in U}\frac{\Delta_{ij}}{\pi_i\pi_j}\left(\sum_{l=1}^{L} w_l(t)Y_i(t_l)\right)\left(\sum_{l'=1}^{L} w_{l'}(t)Y_j(t_{l'})\right).\end{aligned} \tag{5.26}$$

En combinant (5.25) et (5.26) dans (5.24), on obtient après quelques calculs l'expression suivante pour la variance de $\widehat{\mu}_{r,\mathrm{HT}}(t)$, à chaque instant t de $[0, T]$,

$$\mathbb{V}(\widehat{\mu}_{r,\mathrm{HT}}(t)) = \frac{1}{N^2}\sum_{i,j\in U}\frac{\Delta_{ij}}{\pi_i\pi_j}\left(\sum_{l=1}^{L} w_l(t)Y_i(t_l)\right)\left(\sum_{l'=1}^{L} w_{l'}(t)Y_l(t_{l'})\right) \tag{5.27}$$

$$+ \frac{1}{N^2}\sum_{i\in U}\frac{1}{\pi_i}\sum_{l,l'=1}^{L} w_l(t)w_{l'}(t)Y_i(t_l)Y_i(t_{l'})\frac{\vartheta_i(t_l,t_{l'}) - \vartheta_i(t_l)\vartheta_i(t_{l'})}{\vartheta_i(t_l).\vartheta_i(t_{l'})} \tag{5.28}$$

avec la convention $\vartheta_i(t_l, t_l) = \vartheta_i(t_l)$. La part de la variance donnée dans (5.27) correspond à la variance d'échantillonnage tandis que le terme de variance additionnel de (5.28) est dû à la non réponse.

Approximation de variance pour les estimateurs de Hájek

La variance de l'estimateur $\widehat{\mu}_{r,\mathrm{Ha}}^{(1)}(t)$ défini par (5.11) peut être approximée par linéarisation (voir Deville (1999)) sous le plan de sondage et le mécanisme de non réponse. En effet, $\widehat{\mu}_{r,\mathrm{Ha}}^{(1)}(t)$ est un ratio de deux estimateurs linéaires dont les expressions sont similaires à celle de $\widehat{\mu}_{r,\mathrm{HT}}(t)$. On a donc

$$\mathbb{V}\left(\widehat{\mu}_{r,\mathrm{Ha}}^{(1)}(t)\right) \approx \mathbb{V}\left(\sum_{i\in s}\sum_{l=1}^{L}\frac{r_i(t_l)}{\vartheta_i(t_l)}\frac{u_{il}^{(1)}(t)}{\pi_i}\right), \tag{5.29}$$

où la variable linéarisée $u_{il}^{(1)}(t)$ est définie de la manière suivante

$$u_{il}^{(1)}(t) = \frac{1}{\mathrm{N}} w_l(t) \left(\mathrm{Y}_i(t_l) - \tilde{\mu}(t)\right). \tag{5.30}$$

Après quelques calculs, on obtient l'expression suivante de la variance :

$$\mathbb{V}\left(\widehat{\mu}_{r,\mathrm{H}a}^{(1)}(t)\right) \approx \frac{1}{\mathrm{N}^2} \sum_{i\in \mathrm{U}} \sum_{l\in \mathrm{U}} \frac{\Delta_{ij}}{\pi_i \pi_j} \widetilde{u}_i^{(1)}(t) \widetilde{u}_j^{(1)}(t) + \sum_{i\in \mathrm{U}} \frac{1}{\pi_i} \sum_{l,l'=1}^{\mathrm{L}} \frac{u_{il}^{(1)}(t) u_{il'}^{(1)}(t)}{\vartheta_i(t_l)\vartheta_i(t_{l'})} (\vartheta_i(t_l, t_{l'}) - \vartheta_i(t_l)\vartheta_i(t_{l'})), \tag{5.31}$$

où $\widetilde{u}_i^{(1)}(t)$ est, à un facteur multiplicatif N près, la courbe linéarisée lissée,

$$\widetilde{u}_i^{(1)}(t) \;=\; \sum_{l=1}^{\mathrm{L}} w_l(t) \left(\mathrm{Y}_i(t_l) - \tilde{\mu}(t)\right).$$

Pour le troisième estimateur, $\widehat{\mu}_{r,\mathrm{H}a}^{(2)}(t)$ défini dans (5.12), on a,

$$\mathbb{V}\left(\widehat{\mu}_{r,\mathrm{H}a}^{(2)}(t)\right) = \sum_{l,l'=1}^{\mathrm{L}} w_l(t) w_{l'}(t)\, \mathbb{C}ov\left(\frac{\widehat{\mathrm{Y}}(t_l)}{\widehat{\mathrm{N}}(t_l)}, \frac{\widehat{\mathrm{Y}}(t_{l'})}{\widehat{\mathrm{N}}(t_{l'})}\right). \tag{5.32}$$

En utilisant une fois de plus la linéarisation, on obtient que

$$\mathbb{C}ov\left(\frac{\widehat{\mathrm{Y}}(t_l)}{\widehat{\mathrm{N}}(t_l)}, \frac{\widehat{\mathrm{Y}}(t_{l'})}{\widehat{\mathrm{N}}(t_{l'})}\right) \approx \mathbb{C}ov\left(\frac{1}{\mathrm{N}} \sum_{i\in s} \frac{r_i(t_l)}{\vartheta_i(t_l)} \frac{(\mathrm{Y}_i(t_l) - \mu(t_l))}{\pi_i}, \frac{1}{\mathrm{N}} \sum_{i\in s} \frac{r_i(t_{l'})}{\vartheta_i(t_{l'})} \frac{(\mathrm{Y}_i(t_{l'}) - \mu(t_{l'}))}{\pi_i}\right),$$

de sorte que

$$\begin{aligned}\mathbb{V}\left(\widehat{\mu}_{r,\mathrm{H}a}^{(2)}(t)\right) &\approx \mathbb{V}\left(\sum_{l=1}^{\mathrm{L}} w_l(t) \frac{1}{\mathrm{N}} \sum_{i\in s} \frac{r_i(t_l)}{\vartheta_i(t_l)} \frac{(\mathrm{Y}_i(t_l) - \mu(t_l))}{\pi_i}\right) \\ &= \frac{1}{\mathrm{N}^2} \mathbb{V}\left(\sum_{i\in s} \sum_{l=1}^{\mathrm{L}} \frac{r_i(t_l)}{\vartheta_i(t_l)} \frac{w_l(t)(\mathrm{Y}_i(t_l) - \mu(t_l))}{\pi_i}\right).\end{aligned} \tag{5.33}$$

Une comparaison directe de (5.33) avec (5.29) nous montre que l'approximation de variance de $\widehat{\mu}_{r,\mathrm{H}a}^{(2)}(t)$ basée sur la linéarisation est très proche de l'approximation de variance de $\widehat{\mu}_{r,\mathrm{H}a}^{(1)}(t)$. On a

$$\mathbb{V}\left(\widehat{\mu}_{r,\mathrm{H}a}^{(2)}(t)\right) \approx \frac{1}{\mathrm{N}^2} \sum_{i\in \mathrm{U}} \sum_{j\in \mathrm{U}} \frac{\Delta_{ij}}{\pi_i \pi_j} \widetilde{u}_i^{(2)}(t) \widetilde{u}_j^{(2)}(t) \quad + \tag{5.34}$$

$$\sum_{i\in \mathrm{U}} \frac{1}{\pi_i} \sum_{l,l'=1}^{\mathrm{L}} \frac{u_{il}^{(2)}(t) u_{il'}^{(2)}(t)}{\vartheta_i(t_l)\vartheta_i(t_{l'})} (\vartheta_i(t_l, t_{l'}) - \vartheta_i(t_l)\vartheta_i(t_{l'})),$$

où

$$u_{il}^{(2)} \;=\; \frac{1}{\mathrm{N}} w_l(t)(\mathrm{Y}_i(t_l) - \mu(t_l))$$

est la variable linéarisée de $\mu(t_l), l = 1, \ldots, \mathrm{L}$, avec

$$\sum_{l=1}^{\mathrm{L}} \sum_{i\in \mathrm{U}} w_l(t) u_{il}^{(2)} = 0$$

et

$$\widetilde{u}_i^{(2)}(t) = \sum_{l=1}^{\mathrm{L}} w_l(t)\left(\mathrm{Y}_i(t_l) - \mu(t_l)\right).$$

Comme $\sum_{l=1}^{\mathrm{L}} w_l(t) = 1$ on a aussi $\widetilde{u}_l^{(2)}(t) \;=\; \sum_{l=1}^{\mathrm{L}} w_l(t)\left(\mathrm{Y}_i(t_l) - \widetilde{\mu}(t)\right)$.

Estimation de variance pour un estimateur de type Hájek. Pour un estimateur de la moyenne ou du total par lissage de type Hájek, on a besoin d'estimer les valeurs des variables linéarisées afin de construire un estimateur de variance. On peut considérer par exemple l'estimateur de variance suivant (voir Ardilly and Tillé (2006), chapitre 9) :

$$\begin{aligned}\widehat{\mathbb{V}}\left(\widehat{\mu}_{r,\mathrm{H}a}^{(1)}(t)\right) &= \frac{1}{\mathrm{N}^2}\sum_{i\in s}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}\pi_i\pi_j}\widehat{\widetilde{u}}_i(t)\widehat{\widetilde{u}}_j(t)\\ &\quad+\sum_{i\in s}\frac{1}{\pi_i}\sum_{l,l'=1}^{\mathrm{L}}\widehat{u}_{il}(t)\widehat{u}_{il'}(t)\frac{(\vartheta_i(t_l,t_{l'})-\vartheta_i(t_l)\vartheta_i(t_{l'}))}{\vartheta_i(t_l)\vartheta_i(t_{l'})}r_i(t_l)r_i(t_{l'}),\end{aligned}$$

où

$$\widehat{u}_{il}(t) = \frac{1}{\mathrm{N}}w_l(t)\left(\mathrm{Y}_i(t_l) - \widehat{\mu}_{r,\mathrm{H}a}^{(1)}(t)\right)$$

et

$$\widehat{\widetilde{u}}_i(t) \;=\; \sum_{l=1}^{\mathrm{L}} w_l(t)\left(\frac{r_i(t_l)}{\vartheta_i(t_l)}\mathrm{Y}_i(t_l) - \widehat{\mu}_{r,\mathrm{H}a}^{(1)}(t)\right).$$

Formule de variance pour le sondage stratifié

Ici, comme dans le paragraphe 5.2.1, on considère le cas particulier fréquent du sondage stratifié, avec sondage aléatoire simple dans chaque strate. L'expression de la courbe moyenne est donnée par (5.13). On note que par indépendance des échantillons dans chaque strate $s_1,\ldots,s_{\mathrm{H}}$, on a

$$\mathbb{V}(\widehat{\mu}(t)) = \frac{1}{\mathrm{N}^2}\sum_{h=1}^{\mathrm{H}}\mathrm{N}_h^2\,\mathbb{V}(\widehat{\mu}_h(t)). \tag{5.35}$$

En considérant l'estimateur $\widehat{\mu}_{h,\mathrm{HT}}(t)$de $\widetilde{\mu}_h$, on obtient

$$\begin{aligned}\mathbb{V}(\widehat{\mu}_{h,\mathrm{HT}}(t)) &= \left(1-\frac{n_h}{\mathrm{N}_h}\right)\frac{1}{n_h}\frac{1}{\mathrm{N}_h-1}\sum_{i\in\mathrm{U}_h}\left(\widetilde{\mathrm{Y}}_i(t)-\widetilde{\mu}_h(t)\right)^2\\ &\quad+\frac{\mathrm{N}_h}{n_h}\sum_{i\in\mathrm{U}_h}\sum_{l,l'=1}^{\mathrm{L}}w_l(t)w_{l'}(t)\mathrm{Y}_i(t_l)\mathrm{Y}_i(t_{l'})\frac{\vartheta_h(t_l,t_{l'})-\vartheta_h(t_l)\vartheta_h(t_{l'})}{\vartheta_h(t_l)\vartheta_h(t_{l'})},\end{aligned} \tag{5.36}$$

où $\widetilde{\mathrm{Y}}_i(t) = \sum_{l=1}^{\mathrm{L}} w_l(t)\mathrm{Y}_i(t_l)$ est la trajectoire lissée de l'unité i, lorsqu'il n'y a pas de non réponse. Si on considère plutôt un estimateur par le ratio, on a

$$\begin{aligned}\mathbb{V}(\widehat{\mu}_{h,\mathrm{H}a}^{(1)}(t)) &\approx \left(1-\frac{n_h}{\mathrm{N}_h}\right)\frac{1}{n_h}\frac{1}{\mathrm{N}_h-1}\sum_{i\in\mathrm{U}_h}\left(\widetilde{\mathrm{Y}}_i(t)-\widetilde{\mu}_h(t)\right)^2\\ &\quad+\frac{\mathrm{N}_h}{n_h}\sum_{i\in\mathrm{U}_h}\sum_{l,l'=1}^{\mathrm{L}}w_l(t)w_{l'}(t)(\mathrm{Y}_i(t_l)-\widetilde{\mu}_h(t))(\mathrm{Y}_i(t_{l'})-\widetilde{\mu}_h(t))\frac{\vartheta_h(t_l,t_{l'})-\vartheta_h(t_l)\vartheta_h(t_{l'})}{\vartheta_h(t_l)\vartheta_h(t_{l'})}\end{aligned} \tag{5.37}$$

et

$$\mathbb{V}(\widehat{\mu}^{(2)}_{h,\mathrm{H}a}(t)) \approx \left(1-\frac{n_h}{\mathrm{N}_h}\right)\frac{1}{n_h}\frac{1}{\mathrm{N}_h-1}\sum_{i\in \mathrm{U}_h}\left(\widetilde{\mathrm{Y}}_i(t)-\widetilde{\mu}_h(t)\right)^2$$
$$+\frac{\mathrm{N}_h}{n_h}\sum_{i\in \mathrm{U}_h}\sum_{l,l'=1}^{\mathrm{L}} w_l(t)w_{j'}(t)(\mathrm{Y}_i(t_l)-\mu_h(t_l))(\mathrm{Y}_i(t_{l'})-\mu_h(t_l))\frac{\vartheta_h(t_l,t_{l'})-\vartheta_h(t_{l'})\vartheta_h(t_{l'})}{\vartheta_h(t_l)\vartheta_h(t_{l'})}. \tag{5.38}$$

On note que, comme attendu, la part de la variance due à l'erreur d'échantillonnage est la même pour les deux estimateurs car ceux-ci coïncident lorsqu'il n'y a pas de valeurs manquantes.

Comparaison des variances pour le sondage stratifié

Dans cette sous-section, nous comparons les variances des différents estimateurs de courbes moyennes lissées définis dans 5.2.1 pour une strate $h = 1,\dots \mathrm{H}$ donnée. On rappelle qu'on considère ici que les strates sont des groupes de réponse homogènes. Dans cette strate, l'échantillon s_h a été sélectionné par sondage aléatoire simple. On définit

$$\Delta_h(l,l') = \frac{\vartheta_h(t_l,t_{l'})-\vartheta_h(t_l)\vartheta_h(t_{l'})}{\vartheta_h(t_l)\vartheta_h(t_{l'})}.$$

La différence entre les variances des estimateurs peut être approximée par

$$\mathbb{V}(\widehat{\mu}^{(1)}_{h,\mathrm{H}a}(t)) - \mathbb{V}(\widehat{\mu}_{h,\mathrm{HT}}(t)) \approx \frac{\mathrm{N}_h}{n_h}\sum_{l,l'=1}^{\mathrm{L}}\Delta_h(l,l')w_l(t)w_{l'}(t)\sum_{i\in \mathrm{U}_h}\widetilde{\mu}_h(t)(\widetilde{\mu}_h(t)-\mathrm{Y}_i(t'_l)-\mathrm{Y}_i(t_l))$$
$$= \frac{\mathrm{N}_h^2}{n_h}\widetilde{\mu}_h(t)\sum_{l,l'=1}^{\mathrm{L}}\Delta_h(l,l')w_l(t)w_{l'}(t)\left(\widetilde{\mu}_h(t))-\mu_h(t_l)-\mu_h(t'_l)\right)$$

Considérons maintenant la matrice de taille $\mathrm{L}\times\mathrm{L}$ $\mathbf{\Delta}_h$, dont les éléments génériques sont $\Delta_h(l,l')$, la différence ci-dessus s'exprime alors comme

$$\mathbb{V}(\widehat{\mu}^{(1)}_{h,\mathrm{H}a}(t)) - \mathbb{V}(\widehat{\mu}_{h,\mathrm{HT}}(t)) \approx \frac{\mathrm{N}_h^2}{n_h}\left(\mathbf{w}(t)\widetilde{\mu}_h(t)\right)^{\mathrm{T}}\mathbf{\Delta}_h\left(\mathbf{w}(t)\widetilde{\mu}_h(t)-2\breve{\boldsymbol{\mu}}_h(t)\right)$$

avec $\mathbf{w}(t) = (w_1(t),\dots,w_{\mathrm{L}}(t))$ et $\breve{\boldsymbol{\mu}}_h(t) = (w_1(t)\mu_h(t_1),\dots,w_{\mathrm{L}}(t)\mu_h(t_{\mathrm{L}}))$.

Comme la fenêtre η est petite, $w_l(t)$ est très faible (et supposé négligeable) lorsque t n'est pas très proche de t_l et $w_l(t)\approx 1$ si $t\approx t_l$. On peut donc faire l'approximation suivante :

$$\left(\mathbf{w}(t)\widetilde{\mu}_h(t)-2\breve{\boldsymbol{\mu}}_h(t)\right)\approx -\mathbf{w}(t)\widetilde{\mu}_h(t).$$

En remarquant que la matrice $\mathbf{\Delta}_h$ est non négative (c'est une matrice de variance covariance), on obtient finalement que, pour $l = 1,\dots,\mathrm{L}$,

$$\mathbb{V}(\widehat{\mu}^{(1)}_{h,\mathrm{H}a}(t)) - \mathbb{V}(\widehat{\mu}_{\eta,\mathrm{HT}}(t)) \approx -\left(\mathbf{w}(t)\widetilde{\mu}_h(t)\right)^{\mathrm{T}}\mathbf{\Delta}_h\left(\mathbf{w}(t)\widetilde{\mu}_h(t)\right)$$
$$\leq 0$$

et donc que l'estimateur de Hájek $\widehat{\mu}^{(1)}_{h,\mathrm{H}a}(t)$ semble préférable à celui de Horvitz-Thompson $\widehat{\mu}_{h,\mathrm{HT}}(t)$ car sa variance est plus faible lorsque la valeur de la fenêtre η est petite.

5.3 Estimation de courbes moyennes ou totales en présence de valeurs manquantes par imputation

Dans cette section, nous proposons des méthodes d'estimation par imputation. Il s'agit alors de remplacer les valeurs manquantes par des valeurs jugées pertinentes, de façon à limiter la perte de précision due à la présence des valeurs manquantes.

5.3.1 Imputation par des estimateurs à noyau de la courbe de charge

Par construction, l'estimateur à noyau présenté dans la section précédente sera lisse. Or on constate souvent sur nos données que la courbe moyenne que nous cherchons à estimer ne l'est pas. Une idée pour améliorer l'estimation de la courbe moyenne ou totale sera donc de ne plus utiliser directement l'estimateur lisse de la moyenne mais seulement d'imputer les portions de courbes manquantes par l'estimateur lissé et de conserver telles quelles les portions de courbes non manquantes. Notre second estimateur sera donc celui obtenu en imputant par l'estimateur lissé.

Dans chaque classe d'imputation, on va estimer la courbe moyenne à chaque instant par un estimateur à noyau, de la façon décrite au paragraphe précédent. Parmi les différents estimateurs proposés dans la section précédente, on privilégie l'estimateur de Hájek défini par (5.4), dont on montre dans la section 5.2.2 que la variance est plus faible que celle de l'estimateur de Horvitz-Thompson sous des conditions générales et pour le sondage aléatoire simple stratifié. L'estimateur de la moyenne de chaque classe d'imputation U_c s'écrit alors pour tout instant t

$$\widehat{\mu}_c(t) = \sum_{l=1}^{\mathrm{L}} w(t, t_l, \eta) \frac{\displaystyle\sum_{i \in s_c} \frac{r_i(t_l)}{\vartheta_i(t_l)} \frac{\mathrm{Y}_i(t_l)}{\pi_i}}{\displaystyle\sum_{i \in s_c} \frac{r_i(t_l)}{\vartheta_i(t_l)} \frac{1}{\pi_i}}.$$

Dans le cadre de l'étude des consommations électriques, on peut de plus supposer que les classes d'imputation sont confondues avec les groupes de réponse homogènes et que les probabilités de non réponse sont constantes au cours du temps. On a donc $\vartheta_i(t_l) = \vartheta_c \quad \forall i \in \mathrm{U}_c, \quad l = 1, \ldots \mathrm{L}$.

$$\widehat{\mu}_c(t) = \frac{\displaystyle\sum_{\ell=1}^{\mathrm{L}} \mathrm{K}\left(\frac{t - t_\ell}{\eta}\right) \sum_{i \in s_c} r_i(t_\ell) \frac{\mathrm{Y}_i(t_\ell)}{\pi_i}}{\displaystyle\sum_{\ell=1}^{\mathrm{L}} \mathrm{K}\left(\frac{t - t_\ell}{\eta}\right) \sum_{i \in s_c} r_i(t_\ell) \frac{1}{\pi_i}}, \tag{5.39}$$

Pour chaque classe d'imputation, il nous reste alors à remplacer les valeurs manquantes par l'estimateur de la courbe moyenne lissée pour cette classe. Notre estimateur de la courbe totale en présence de valeurs manquantes s'écrit finalement

$$\widehat{t}_{\mathrm{K}}(t) = \sum_{i \in s_r^{(c)}(t)} \frac{\mathrm{Y}_i(t)}{\pi_i} + \sum_{c=1}^{\mathrm{C}} \sum_{i \in s_m^{(c)}(t)} \frac{\widehat{\mu}_c(t)}{\pi_i}. \tag{5.40}$$

5.3.2 Imputation par les plus proches voisins

Comme l'expliquent Chen and Shao (2000) et Haziza (2009), l'imputation par le ou les plus proches voisins a de nombreux avantages qui font que cette méthode est très largement utilisée dans les organismes de statistiques. Il s'agit d'une technique d'imputation par donneur(s) simple à mettre en œuvre, qui consiste à remplacer la valeur manquante par la valeur moyenne des k_n unités observées les plus proches, appelées voisins. Ces voisins sont les k_n unités observées dont les valeurs pour les variables auxiliaires sont les plus proches possibles des variables auxiliaires de l'unité non observée. Tout l'enjeu est donc de trouver les variables auxiliaires, et donc les voisins, les plus pertinents possibles ainsi qu'une distance pertinente pour quantifier la proximité.

L'imputation par les plus proches voisins n'impose pas de forme paramétrique rigide pour la relation entre la variable d'intérêt et les variables auxiliaires pertinentes. Cependant, il faut veiller à ne considérer qu'un nombre faible de variables auxiliaires. En effet, un des inconvénients de cette méthode inhérent à son caractère non paramétrique est que le biais peut être non négligeable par rapport à la variance lorsque le nombre de variables auxiliaires considérées n'est pas petit. On peut en effet montrer dans le cadre fréquentiste classique que le biais est d'ordre $(k_n/n)^{1/p}$ où k_n est le nombre de voisins, n la taille de l'échantillon et p le nombre de variables auxiliaires (Györfi et al. (2002), chapitre 6).

Lorsque notre objectif est de remplacer les valeurs manquantes d'une courbe incomplète (on est alors dans la première problématique évoquée dans l'introduction et que nous ne traitons pas ici), le compromis entre biais et variance conduit en général à sélectionner un nombre de voisins réduit mais cependant supérieur à 1 (ce nombre pouvant être choisi par validation croisée). Cependant dans notre contexte où l'objectif final est d'estimer le plus précisément possible la courbe moyenne ou totale comme dans (5.3), on sélectionne en général un plus petit nombre de voisins que précédemment, afin de garder le biais aussi faible que possible. On note en outre que les variables explicatives doivent être sélectionnées soigneusement afin que p soit aussi faible que possible.

Lorsqu'une courbe Y_i est partiellement observée, la spécificité de notre problématique fonctionnelle est que l'on cherche un ou plusieurs donneurs dont la trajectoire sur les instants observés pour i soit aussi proche que possible des portions observées de Y_i. Ainsi, pour une trajectoire Y_i inobservée aux $l' - l + 1$ instants, $t_l, t_{l+1}, \ldots, t_{l'}$, on peut donc mesurer sa proximité par rapport aux unités complètement observées appartenant à la même classe d'imputation en prenant en compte l'ensemble des $\ell = \mathrm{L} - (l' - l + 1)$ points de mesure correspondant aux valeurs observées de la courbe. Cela nous conduirait cependant à considérer un nombre de variables explicatives ℓ potentiellement élevé, ce qui n'est pas souhaitable.

Du fait des très fortes corrélations temporelles présentes dans notre contexte fonctionnel, nous avons choisi de ne retenir comme variables explicatives que les valeurs de Y aux extrémités de la partie non observée de la courbe t_{l-1} et $t_{l'+1}$. Avec ce choix, le nombre de variables auxiliaires considérées $\ell = 2$, reste faible, ce qui permet de garder un biais modéré. Par conséquent, l'information auxiliaire utilisée pour choisir le donneur dépendra de chaque cas particulier car t_{l-1} et $t_{l'+1}$ varient d'une courbe partiellement observée à une autre et même pour plusieurs séquences manquantes d'une même trajectoire. Dans ce cadre général, cela signifie donc que pour chaque séquence

de valeurs manquantes, on considère implicitement un modèle d'imputation différent.

Le fait d'utiliser des classes d'imputation permet d'améliorer la qualité de l'estimation en évitant de choisir un donneur dont la trajectoire est très différente de celle de la courbe à imputer à part pour les points t_{l-1} et $t_{l'+1}$. On exploite donc simultanément l'information de "court terme", c'est-à-dire les valeurs de la courbe juste avant et juste après la séquence manquante et l'information de "long terme" qui peut regrouper des informations auxiliaires plus générales. En outre, l'usage de classes d'imputation, en limitant le nombre de voisins potentiels, permet de conserver des temps de calcul raisonnables même pour des périodes d'études longues et de gros échantillons : il suffit alors de découper la population en classes d'imputation plus fines.

Pour une courbe à imputer Y_i appartenant à la classe U_c, les plus proches voisins sont donc recherchés parmi les individus (les courbes) appartenant à la même classe. Ainsi, soit Y_i une courbe inobservée entre t_l et $t_{l'}$ inclus, on cherche des unités $i' \in s_c$ telles que

$$(\mathrm{Y}_i(t_{l-1}) - \mathrm{Y}_{i'}(t_{l-1}))^2 + (\mathrm{Y}_i(t_{l'+1}) - \mathrm{Y}_{i'}(t_{l'+1}))^2$$

soit minimale parmi les unités de s_c sous la contrainte que la courbe Y'_i soit complètement observée entre t_l et $t_{l'}$. De plus, cette procédure d'imputation permet de garantir la cohérence temporelle de la courbe imputée car l'intégralité de la séquence imputée entre t_{l-1} et $t_{l'+1}$ provient du ou des mêmes donneurs. La procédure de sélection du donneur est illustrée par la Figure 5.1 dans le cas d'un donneur unique.

Notons qu'il serait également possible de prendre en compte la pente des trajectoires incomplètes juste avant et juste après la séquence manquante en considérant en tant que variables explicatives additionnelles les différences $\mathrm{Y}_i(t_{l-1}) - \mathrm{Y}_i(t_{l-2})$ et $\mathrm{Y}_i(t_{l'+2}) - \mathrm{Y}_i(t_{l'+1})$. Cela revient à ajouter deux variables explicatives supplémentaires, et n'améliore pas la qualité des estimations dans nos simulations.

Finalement, notre estimateur des plus proches voisins peut s'écrire, pour $t \in \{t_1, \ldots, t_{\mathrm{L}}\}$,

$$\hat{t}_{nn}(t) = \sum_{i \in s_r(t)} \frac{\mathrm{Y}_i(t)}{\pi_i} + \sum_{c=1}^{\mathrm{C}} \sum_{i \in s_m^{(c)}(t)} \frac{\mathrm{Y}_i^*(t)}{\pi_i} \tag{5.41}$$

avec $\mathrm{Y}_i^*(t) = \mathrm{Y}_{j_{c(i)}}(t)$ la valeur à l'instant t de la courbe du donneur $j_{c(i)} \in s_r^{(c)}(t) = s_r(t) \cap \mathrm{U}_c$ pour l'unité manquante $i \in s_m^{(c)}(t) = s_m(t) \cap \mathrm{U}_c$. Lorsque l'on considère un nombre de donneurs k strictement supérieur à 1, $\mathrm{Y}_{j_c(i)}(t)$ est remplacé par la portion de courbe moyenne des k-plus proches voisins dans la classe d'imputation U_c.

5.3.3 Interpolation linéaire de la différence à la moyenne

L'une des méthodes d'imputation la plus simple à mettre en œuvre pour des courbes partiellement observées est l'interpolation linéaire. Elle consiste simplement à remplacer les valeurs manquantes en réalisant une interpolation linéaire entre le premier et le dernier point observés pour chaque séquence manquante. Pour une trajectoire observée en t_{l-1} et $t_{l'+1}$ mais manquante entre t_l et $t_{l'}$ inclus, on utilise le processus d'imputation

$$\mathrm{Y}_i^*(t) = \frac{t_{l'+1} - t}{t_{l'+1} - t_{l-1}} \mathrm{Y}_i(t_{l-1}) + \frac{t - t_{l-1}}{t_{l'+1} - t_{l-1}} \mathrm{Y}_i(t_{l'+1}), \quad t \in \{t_d, \ldots, t_{d'}\}. \tag{5.42}$$

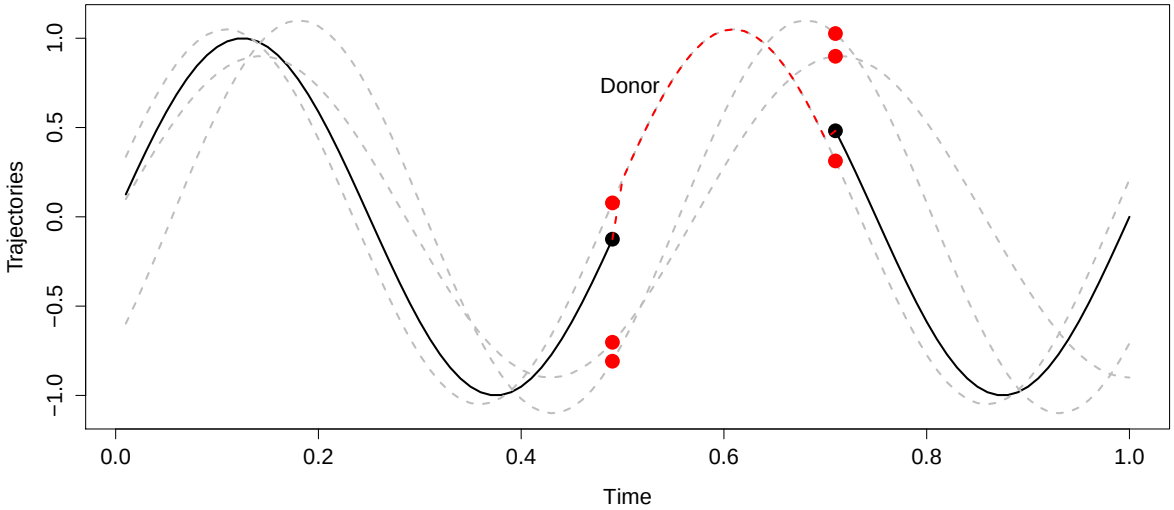

FIGURE 5.1 – Imputation par le plus proche voisin pour la courbe partiellement inobservée en noir. Parmi les trois donneurs potentiels (courbes en pointillés gris), le plus proche voisin sélectionné pour l'imputation est la courbe la plus proche au début et à la fin de la séquence manquante (points noirs et rouges). Les valeurs imputées sont en pointillés rouges.

Lorsque la séquence manquante est longue, l'interpolation linéaire ne permet pas de prendre en compte la tendance commune à l'ensemble des courbes de la population, qui peut potentiellement être non linéaire. On va donc utiliser une variante de cette méthode pour pallier ce défaut. Supposons que l'on réalise maintenant l'interpolation linéaire sur les variations autour de la courbe moyenne de la classe d'imputation U_c, pour $t \in \{t_d, \ldots, t_{d'}\}$,

$$\mathrm{Y}_i^*(t) = \hat{\mu}_c(t) + \frac{t_{l'+1} - t}{t_{l'+1} - t_{l-1}} \left(\mathrm{Y}_i(t_{l-1}) - \hat{\mu}_c(t_{l-1})\right) + \frac{t - t_{l-1}}{t_{l'+1} - t_{l-1}} \left(\mathrm{Y}_i(t_{l'+1}) - \hat{\mu}_c(t_{l'+1})\right), \tag{5.43}$$

avec $\hat{\mu}_c(t) = \left(\sum_{s_c} d_i r_i(t) \mathrm{Y}_i(t)\right) / \left(\sum_{s_c} d_i r_i(t)\right)$ la courbe moyenne estimée (avec des pondérations de type Hájek) parmi les répondants de la classe U_c. En suivant (5.43), les valeurs imputées auront la même forme commune que les unités observées de leur classe d'imputation. Notons que lorsque la courbe est intégralement manquante, on l'impute par la moyenne de classe. Les deux procédures d'interpolations linéaires (simples et de la différence à la moyenne) sont illustrées dans la Figure 5.2.

Après imputation, on en déduit simplement l'estimateur du total :

$$\hat{t}_{int}(t) = \sum_{i \in s_r(t)} \frac{\mathrm{Y}_i(t)}{\pi_i} + \sum_{c=1}^{\mathrm{C}} \sum_{i \in s_m^{(c)}(t)} \frac{\mathrm{Y}_i^*(t)}{\pi_i}, \quad t \in \{t_1, \ldots, t_{\mathrm{L}}\}. \tag{5.44}$$

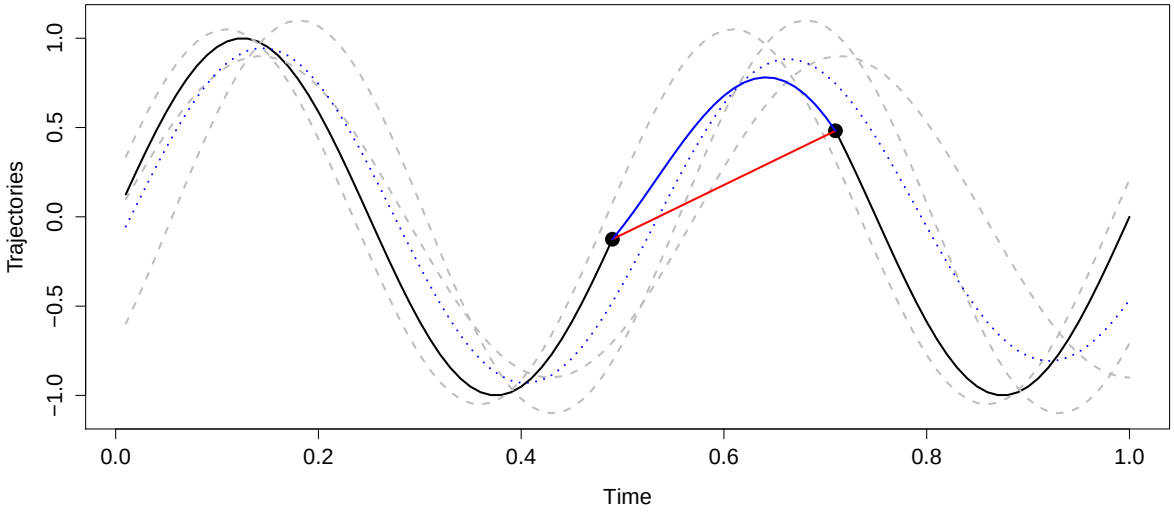

FIGURE 5.2 – Imputation basée sur l'interpolation linéaire pour la courbe partiellement observée en noir. La courbe moyenne des unités répondantes au sein de la classe d'imputation est tracée en pointillés bleus. L'imputation par interpolation linéaire est en rouge (traits continus) et l'interpolation des différences à la moyenne de classe est en bleu (traits continus).

5.3.4 Approche unifiée pour l'approximation de variance des estimateurs par imputation

Dans cette section, nous proposons des estimateurs de variance pour chacune des méthodes d'imputation proposées dans les sous-sections 5.3.1, 5.3.2 et 5.3.3. On se place dans le cadre général des estimateurs linéaires basés sur l'imputation linéaire. Ce cadre de travail est relativement similaire à celui introduit par Beaumont and Bissonnette (2011) dans le contexte des modèles d'imputation composite [3]. Notre contexte est cependant un peu différent puisque la valeur imputée à un instant t peut dépendre linéairement de valeurs observées à d'autres instants, avec un ensemble des répondants qui varie au cours du temps.

Les estimateurs par imputation proposés dans les sections précédentes, que ce soit l'imputation par lissage à noyau, l'imputation par les plus proches voisins ou l'interpolation linéaire peuvent être vus comme des cas particuliers de ce cadre général. On considère ici le cas de l'estimation d'une courbe totale, mais l'adaptation au cas de l'estimation d'une courbe moyenne est triviale.

3. Il s'agit de procédures d'imputation dans lesquelles l'information auxiliaire, et donc les modèles employés varient d'une unité à l'autre.

Imputation linéaire

En utilisant les mêmes notations que dans Beaumont and Bissonnette (2011), l'ensemble de nos estimateurs par imputation du total t_Y à l'instant t peuvent s'écrire

$$\hat{t}_{Y,imp}(t) = \sum_{i \in s_r(t)} \frac{Y_i(t)}{\pi_i} + \sum_{c=1}^{C} \sum_{i \in s_m^{(c)}(t)} \frac{Y_i^*(t)}{\pi_i} \tag{5.45}$$

où $Y_i^*(t)$ est la valeur imputée pour $Y_i(t)$ manquante. Lorsque l'unité i dont la courbe est incomplète appartient à la classe U_c, sa valeur imputée $Y_i^*(t)$ peut s'écrire comme une combinaison linéaire de trajectoires observées non seulement à l'instant t, mais aussi avant et après t dans la classe d'imputation U_c. On considère donc que les valeurs imputées peuvent s'écrire de la manière suivante en fonction des observations de la même classe d'imputation,

$$Y_i^*(t) = \varphi_{0i}^{(c)} + \sum_{\ell=1}^{L} \sum_{j \in s_r^{(c)}(t_\ell)} \varphi_{ji}^{(c)}(t, t_\ell) Y_j(t_\ell). \tag{5.46}$$

Dans le cas particulier de l'imputation par les $k-$plus proches voisins, $\varphi_{0i}^{(c)} = 0$, $\varphi_{ji}^{(c)}(t, t_\ell) = \frac{1}{k}, \quad i \in s_c$, si j fait partie des k plus proches voisins de i dans la classe c et $t_\ell = t$, $\varphi_{ji}^{(c)}(t, t_\ell) = 0$ si j n'est pas un voisin ou $t_\ell \neq t$. L'instant t est le seul considéré et la valeur imputée ne dépend que des données concernant un seul instant :

$$Y_i^*(t) = \sum_{j \in s_r^{(c)}(t)} \varphi_{ji}^{(c)}(t, t) Y_j(t). \tag{5.47}$$

Pour l'imputation basée sur l'interpolation linéaire, les valeurs des autres unités observées ne sont pas prises en compte et, comme dans (5.42), on a pour $t \in \{t_l, \ldots, t_{l'}\}$,

$$\begin{aligned} Y_i^*(t) &= \varphi_{ii}^{(c)}(t, t_{l-1}) Y_i(t_{l-1}) + \varphi_{ii}^{(c)}(t, t_{l'+1}) Y_i(t_{l'+1}) \\ &= \frac{t_{l'+1} - t}{t_{l'+1} - t_{l-1}} Y_i(t_{l-1}) + \frac{t - t_{l-1}}{t_{l'+1} - t_{l-1}} Y_i(t_{l'+1}), \end{aligned}$$

Les poids dépendent uniquement des parties observées de la trajectoire. De plus, on a $\varphi_{0i}^{(c)} = 0, \quad \forall i \in s$.

Pour l'imputation par le lissage à noyau, l'expression des poids est un peu plus compliquée car le lissage peut potentiellement prendre en compte l'ensemble des données de la classe d'imputation considérée. Pour une unité i appartenant à la classe U_c, les poids ne varient pas en fonction de i et ne dépendent que de la classe c à laquelle l'unité appartient.

$$\varphi_{ji}^{(c)}(t, t_\ell) = \left(\sum_{\ell=1}^{L} K\left(\frac{t - t_\ell}{\eta}\right) \sum_{i \in s_r^{(c)}(t_\ell)} \frac{1}{\pi_i} \right)^{-1} \frac{K\left(\frac{t - t_\ell}{\eta}\right)}{\pi_j} r_j(t_\ell),$$

et $\varphi_{0i}^{(c)} = 0$. Pour finir, on remarque que les valeurs imputées $Y_i^*(t)$ obtenues dans (5.43) par interpolation linéaire des variations autour de la moyenne sont une combinaison linéaire $Y_i^*(t)$ et de l'estimateur de la moyenne. Cet estimateur peut donc être exprimé comme dans (5.46).

Remarque 10. *En pratique, l'interpolation linéaire n'est pas forcément utilisée de manière exclusive : en effet, on la combine fréquemment avec l'imputation à la moyenne de classe : les séquences trop longues (selon un seuil à définir) ou encore survenant en début ou fin de période temporelle sont en effet imputées par moyenne de classe plutôt que par interpolation linéaire. Cependant, la moyenne de classe peut également s'exprimer comme un estimateur du plus proche voisin (toutes les unités de la classe étant alors des voisins) et donc la combinaison de ces deux méthodes peut également s'exprimer comme dans (5.46).*

Modèle de superpopulation et décomposition générale de la variance.

Pour chaque unité $i \in \mathrm{U}_c$, on suppose que le modèle ci-dessous est vérifié. Pour t et v dans $[0, \mathrm{T}]$ on suppose que

$$\begin{aligned}\mathbb{E}_m(\mathrm{Y}_i(t)) &= \mu_i(t)\\ \mathbb{V}_m(\mathrm{Y}_i(t)) &= \sigma_i^2(t)\\ \mathbb{C}ov_m(\mathrm{Y}_i(t), \mathrm{Y}_j(v)) &= \sigma_i^2(t,v)\mathbb{1}_{\{i=j\}}.\end{aligned}$$

Cette formulation générale, qui sera simplifiée par la suite pour l'interpolation linéaire et le lissage à noyau, suppose que deux trajectoires ne sont pas corrélées mais qu'il peut exister des corrélations temporelles pour chaque trajectoire. En outre, la moyenne $\mu_i(t)$ et la variance $\sigma_i^2(t)$, peuvent varier au cours du temps t et d'une unité i à l'autre.

Pour l'ensemble des estimateurs d'imputation, on supposera (comme dans Beaumont and Bocci (2009) et Beaumont and Bissonnette (2011)) que le biais d'imputation est négligeable. Soient $\mathbb{E}_m$ l'espérance sous le modèle de superpopulation, $\mathbb{E}_p$ l'espérance sous le plan de sondage et $\mathbb{E}_\mathrm{R}$ l'espérance sous le mécanisme de non réponse.Ce biais de l'estimateur imputé est défini de manière générale par

$$\mathbb{E}_m\left(\widehat{t}_{\mathrm{Y},imp}(t) - \widehat{t}_\mathrm{Y}(t) | s, s_r\right).$$

Ici, il est de la forme,

$$\sum_{c=1}^{\mathrm{C}} \sum_{i \in s_m^{(c)}(t)} \frac{1}{\pi_i} \mathbb{E}_m\left(\mathrm{Y}_i(t) - \mathrm{Y}_i^*(t)\right).$$

De plus, on supposera également que le nombre de points de discrétisation est suffisant et les trajectoires assez régulières pour respecter des conditions telles que décrites dans Cardot et al. (2013a) de façon à ce que l'erreur d'approximation due au lissage ou à l'interpolation pour l'estimation à tout instant $t \in [0, \mathrm{T}]$ différent des points de discrétisation soit négligeable par rapport à l'erreur d'échantillonnage. L'erreur quadratique moyenne des totaux sera donc approximativement égale à leur variance.

Si le biais d'imputation est négligeable, alors pour un instant donné t, la variance d'un estimateur basé sur l'imputation $\widehat{t}_{imp}(t)$ peut être approximée par :

$$\mathbb{V}_{tot}(\widehat{t}_{imp}(t)) = \mathbb{E}_m \mathbb{E}_p \mathbb{E}_\mathrm{R}\left(\widehat{t}_{imp}(t) - t_\mathrm{Y}(t)\right)^2, \tag{5.48}$$

On a

$$\widehat{t}_{imp}(t) - t_\mathrm{Y}(t) = \left(\widehat{t}_{imp}(t) - \widehat{t}_\mathrm{Y}(t)\right) + \left(\widehat{t}_\mathrm{Y}(t) - t_\mathrm{Y}(t)\right) \tag{5.49}$$

et donc la variance $\mathbb{V}_{tot}(\widehat{t}_{imp}(t))$ peut se décomposer en

$$\mathbb{V}_{tot}(\widehat{t}_{imp}(t)) = \mathbb{V}_{\mathrm{SAM}}(t) + \mathbb{V}_{\mathrm{IMP}}(t) + 2\mathbb{E}_p\mathbb{E}_{\mathrm{R}}\mathbb{E}_m\left[\left(\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t)\right)\left(\widehat{t}_{\mathrm{Y}}(t) - t_{\mathrm{Y}}(t)\right)|s, s_r\right], \tag{5.50}$$

avec $\mathbb{V}_{\mathrm{SAM}}(t) = \mathbb{E}_m\mathbb{V}_p\left(\widehat{t}_{\mathrm{Y}}(t) - t_{\mathrm{Y}}(t)\right)$ la variance de l'estimateur de Horvitz-Thompson du total $\widehat{t}_{\mathrm{Y}}(t)$ quand il n'y a pas de non réponse et $\mathbb{V}_{\mathrm{IMP}}(t) = \mathbb{E}_p\mathbb{E}_{\mathrm{R}}\mathbb{E}_m\left(\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t)|s, s_r\right)^2$ la variance additionnelle due à l'imputation.

Enfin, le dernier terme de (5.50), est un terme de covariance appelé variance mixte et est noté $\mathbb{V}_{\mathrm{MIX}}(t)$. Il n'est pas toujours négligeable, même lorsque le biais l'est.

Approximation de la variance d'échantillonnage $\mathbb{V}_{\mathrm{SAM}}(t)$ Pour $t \in [0, \mathrm{T}]$, On note $\widehat{v}(t)$ l'estimateur de Horvitz-Thompson de la variance de $\widehat{t}_{\mathrm{Y}}(t)$ lorsqu'il n'y a pas de non réponse. On a

$$\widehat{v}(t) = \sum_{i,j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\mathrm{Y}_i(t)}{\pi_i} \frac{\mathrm{Y}_j(t)}{\pi_j},$$

avec $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$ pour $i \neq j$ et $\Delta_{ii} = \pi_i(1-\pi_i)$. Considérons, comme dans Beaumont and Bocci (2009), l'approximation suivante,

$$\widetilde{\mathbb{V}}_{\mathrm{SAM}}(t) = \mathbb{E}_m\left(\widehat{v}(t)\,|s, s_r, \mathrm{Y}_r\right)$$

avec Y_r la partie observée de Y. Par quelques calculs, on a

$$\widetilde{\mathbb{V}}_{\mathrm{SAM}}(t) = \sum_{k,l\in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\tilde{\mathrm{Y}}_i(t)}{\pi_i} \frac{\tilde{\mathrm{Y}}_j(t)}{\pi_j} + \sum_{c=1}^{\mathrm{C}} \sum_{i\in s_m^{(c)}(t)} \frac{1-\pi_i}{\pi_i^2}\sigma_i^2(t), \tag{5.51}$$

avec $\tilde{\mathrm{Y}}_i(t) = \mathrm{Y}_i(t)$ si $i \in s_r(t)$ et $\tilde{\mathrm{Y}}_i(t) = \mathrm{Y}_i^*(t)$ si $i \in s_m(t)$.

Approximation de la variance d'imputation $\mathbb{V}_{\mathrm{IMP}}(t)$ Si le biais conditionnel peut être considéré comme négligeable, que le plan de sondage est non informatif et que le mécanisme de réponse est ignorable, alors on peut faire l'approximation suivante :

$$\mathbb{V}_{\mathrm{IMP}}(t) \approx \mathbb{E}_p\mathbb{E}_{\mathrm{R}}\mathbb{V}_m\left(\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t)|s, s_r\right). \tag{5.52}$$

En utilisant (5.46), on peut écrire $\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t)$ de la manière suivante :

$$\begin{aligned}
\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t) &= \sum_{i\in s_m(t)} \left(\frac{\mathrm{Y}_i^*(t)}{\pi_i} - \frac{\mathrm{Y}_i(t)}{\pi_i}\right) \\
&= \sum_{c=1}^{\mathrm{C}} \sum_{i\in s_m(t)} \sum_{\ell=1}^{\mathrm{L}} \sum_{j\in s_r^{(c)}(t_\ell)} \frac{\varphi_{ji}^{(c)}(t, t_\ell)}{\pi_i} \mathrm{Y}_l(t_\ell) - \sum_{i\in s_m(t)} \frac{\mathrm{Y}_i(t)}{\pi_i} \\
&= \sum_{c=1}^{\mathrm{C}} \sum_{\ell=1}^{\mathrm{L}} \sum_{j\in s_r^{(c)}(t_\ell)} w_j^{(c)}(t, t_\ell)\mathrm{Y}_l(t_\ell) - \sum_{c=1}^{\mathrm{C}} \sum_{i\in s_m^{(c)}(t)} \frac{\mathrm{Y}_i(t)}{\pi_i},
\end{aligned} \tag{5.53}$$

avec

$$w_j^{(c)}(t,t_\ell) = \sum_{i\in s_m(t)} \frac{\varphi_{ji}^{(c)}(t,t_\ell)}{\pi_i}.$$

Par indépendance des trajectoires dans le modèle de superpopulation, on obtient

$$\mathbb{V}_m\left(\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t)|s,s_r\right) = \sum_{c=1}^{\mathrm{C}} \mathbb{V}_m\left(\sum_{\ell=1}^{\mathrm{L}} \sum_{j\in s_r^{(c)}(t_\ell)} w_j^{(c)}(t,t_\ell)\mathrm{Y}_l(t_\ell) - \sum_{i\in s_m^{(c)}(t)} \frac{\mathrm{Y}_i(t)}{\pi_i}|s,s_r\right). \tag{5.54}$$

En introduisant une fonction indicatrice de réponse à l'instant t_ℓ, $r_i(t_\ell) = 1$ si $i \in s_r(t_\ell)$ et $r_i(t_\ell) = 0$ if $i \in s_m(t_\ell)$, on obtient pour tout $t \in \{t_1,\ldots,t_{\mathrm{L}}\}$, par indépendance entre les trajectoires

$$\begin{aligned}\mathbb{V}_m\left(\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t)|s,s_r\right) &= \sum_{c=1}^{\mathrm{C}} \sum_{i\in s^{(c)}} \mathbb{V}_m\left(\sum_{\ell=1}^{\mathrm{L}} \widetilde{w}_{k,\ell}(t)\mathrm{Y}_i(t_\ell)|s,s_r\right)\\ &= \sum_{c=1}^{\mathrm{C}} \sum_{i\in s^{(c)}} \sum_{\ell,\ell'=1}^{\mathrm{L}} \widetilde{w}_{i,\ell}(t)\widetilde{w}_{i,\ell'}(t)\sigma_i^2(t_\ell,t_{\ell'}),\end{aligned} \tag{5.55}$$

avec

$$\widetilde{w}_{i,\ell}(t) = w_i^{(c)}(t,t_\ell)r_i(t_\ell) - \frac{1}{\pi_i}(1-r_i(t_\ell))\mathbb{1}_{\{t=t_\ell\}}. \tag{5.56}$$

Approximation de la variance mixte $\mathbb{V}_{\mathrm{MIX}}(t)$ En suivant la même approche que pour la variance d'imputation, on obtient

$$\begin{aligned}\mathbb{V}_{\mathrm{MIX}}(t) &= 2\mathbb{E}_p\mathbb{E}_{\mathrm{R}}\mathbb{E}_m\left[\left(\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t)\right)\left(\widehat{t}_{\mathrm{Y}}(t) - t_{\mathrm{Y}}(t)\right)|s,s_r\right]\\ &= 2\mathbb{E}_p\mathbb{E}_{\mathrm{R}}\mathbb{E}_m\left[\left(\sum_{i\in s}\sum_{\ell\in 1}^{\mathrm{L}} \widetilde{w}_{i,\ell}(t)\mathrm{Y}_i(t_\ell)\right)\left(\sum_{i\in s}\left(\frac{1}{\pi_i}-1\right)\mathrm{Y}_i(t)\right)|s,s_r\right]\end{aligned} \tag{5.57}$$

et

$$\mathbb{E}_m\left[\left(\widehat{t}_{imp}(t) - \widehat{t}_{\mathrm{Y}}(t)\right)\left(\widehat{t}_{\mathrm{Y}}(t) - t_{\mathrm{Y}}(t)\right)|s,s_r\right] = \sum_{i\in s}\frac{1-\pi_i}{\pi_i}\left(\sum_{\ell=1}^{p} \widetilde{w}_{i,\ell}(t)\sigma_i^2(t,t_\ell)\right). \tag{5.58}$$

Estimation de variance pour l'imputation par les plus proches voisins

On s'intéresse maintenant au cas particulier de l'imputation par les plus proches voisins. Pour chaque unité manquante $i \in \mathrm{U}_c$ à l'instant $t \in [0,\mathrm{T}]$, la valeur imputée $\mathrm{Y}_i^*(t)$ est une combinaison linéaire des trajectoires observées à l'instant t dans la classe d'imputation U_c,

$$\mathrm{Y}_i^*(t) = \sum_{j\in s_r(t)} \varphi_{ji}^{(c)}\mathrm{Y}_j(t). \tag{5.59}$$

Pour une imputation par les k plus proches voisins, $\varphi_{0i}^{(c)} = 0$ et $\varphi_{j_i i}^{(c)} = 1/k$, avec j_i un des k plus proches voisins et $\varphi_{ji}^{(c)} = 0$ pour $j \neq j_i$.

Pour un instant t, la variance de l'estimateur des plus proches voisins peut donc s'écrire suivant l'équation de décomposition de la variance (5.50). Afin de produire des estimateurs des différentes composantes de cette variance, nous avons besoin d'un estimateur de la variance conditionnelle $\sigma_i^2(t)$. Pour cela, différentes stratégies existent dans la littérature. Par exemple, Beaumont and Bocci (2009) suggèrent de considérer des estimateurs non paramétriques basés sur des splines, avec le risque d'obtenir des estimateurs non positifs. On préfère appliquer l'approche proposée par Shao (2009) qui est simple et efficace. Si $i \in s_r(t) \cap \mathrm{U}_c$, alors

$$\widehat{\sigma}_i^2 = \frac{(\mathrm{Y}_i(t) - \mathrm{Y}_{i^*}(t))^2}{2}, \tag{5.60}$$

avec i^* le plus proche voisin de l'unité i dans $(s_r(t) \cap \mathrm{U}_c) \setminus \{k\}$. Si $i \in s_m(t)$, alors

$$\widehat{\sigma}_c^2 = \frac{(\mathrm{Y}_{i^*}(t) - \mathrm{Y}_{i^{**}}(t))^2}{2}, \tag{5.61}$$

avec i^* et i^{**} les deux plus proches voisins de i appartenant à $s_r(t) \cap \mathrm{U}_c$.

En combinant (5.51) et (5.61), un estimateur de $\widetilde{\mathbb{V}}_{\mathrm{SAM}}(t)$ est

$$\widehat{\mathbb{V}}_{\mathrm{SAM}}(t) = \sum_{i,j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\tilde{\mathrm{Y}}_i(t)}{\pi_i} \frac{\tilde{\mathrm{Y}}_j(t)}{\pi_j} + \sum_{c=1}^{\mathrm{C}} \sum_{i \in s_m^{(c)}(t)} \frac{1-\pi_i}{\pi_i^2} \widehat{\sigma}_c^2(t). \tag{5.62}$$

Lorsque le nombre de voisins k est strictement supérieur à 1, on adapte la méthode en prenant la moyenne des courbes des k plus proches voisins pour les répondants, ou l'écart entre le plus proche voisin et la moyenne des k voisins suivants.

Estimation de la variance d'imputation $\mathbb{V}_{\mathrm{IMP}}(t)$ En utilisant (5.46), on écrit $\widehat{t}_{nn}(t) - \widehat{t}_{\mathrm{Y}}(t)$ de la manière suivante

$$\widehat{t}_{nn}(t) - \widehat{t}_{\mathrm{Y}}(t) = \sum_{i \in s_r(t)} \mathrm{W}_i^{(+)}(t)\mathrm{Y}_i(t) - \sum_{i \in s_m(t)} \frac{\mathrm{Y}_i(t)}{\pi_i}, \tag{5.63}$$

avec

$$\mathrm{W}_i^{(+)}(t) = \sum_{c=1}^{\mathrm{C}} w_i^{(c)}(t) \quad \text{et} \quad w_i^{(c)} = \sum_{j \in s_m^{(c)}(t)} \frac{\varphi_{ji}^{(c)}}{\pi_j}.$$

Le cœfficient $\mathrm{W}_i^{(+)}(t)$ représente le poids additionnel donné à l'observation $\mathrm{Y}_i(t)$ du fait de l'imputation. La variance d'imputation conditionnelle peut être estimée de la manière suivante,

$$\widehat{\mathbb{V}}_{\mathrm{IMP}}\left(\widehat{t}_{nn}(t) - \widehat{t}_{\mathrm{Y}}(t) | s, s_r\right) = \sum_{i \in s_r(t)} \left(\mathrm{W}_i^{(+)}(t)\right)^2 \widehat{\sigma}_i^2(t) + \sum_{i \in s_m(t)} \frac{1}{\pi_i^2} \widehat{\sigma}_i^2(t). \tag{5.64}$$

Estimation de la variance mixte La variance mixte peut aussi être estimée simplement, sous l'hypothèse que le biais est négligeable et on a, comme dans Beaumont and Bissonnette (2011) (equation 5.12),

$$\mathbb{V}_{\text{MIX}}(t) = 2\sum_{i\in s_r(t)} \mathrm{W}_i^{(+)}(t)\frac{1-\pi_i}{\pi_i}\sigma_i^2(t) - 2\sum_{i\in s_m(t)} \frac{1-\pi_i}{\pi_i^2}\sigma_i^2(t). \tag{5.65}$$

Ici encore, ce terme de covariance peut être estimé en remplaçant les $\sigma_i^2(t)$ par leurs estimations définies dans (5.60) et (5.61). Finalement, l'estimateur de variance du plus proche voisin est obtenu en remplaçant (5.62), (5.64) et (5.65) dans la formule globale (5.50).

Estimation de variance pour l'imputation basée sur l'interpolation linéaire

Pour l'interpolation linéaire, la valeur imputée pour une unité non observée i à l'instant t ne dépend que des valeurs observées immédiatement avant et immédiatement après la séquence manquante. Par souci de simplification, on suppose que les termes de variance $\sigma_i^2(t)$ et de covariance $\sigma_i^2(t,v)$ dans les modèles de superpopulation ne varient pas en fonction des unités i dans chaque classe d'imputation U_c :
Si $i \in \mathrm{U}_c$, $\sigma_i^2(t) = \sigma_c^2(t)$ et $\sigma_i^2(t,v) = \sigma_c^2(t,v)$.

La variance d'échantillonnage $\mathbb{V}_{\text{SAM}}(t)$ peut être estimée par

$$\widehat{\mathbb{V}}_{\text{SAM}}(t) = \sum_{i,j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\frac{\tilde{\mathrm{Y}}_i(t)}{\pi_i}\frac{\tilde{\mathrm{Y}}_j(t)}{\pi_j} + \sum_{c=1}^{\mathrm{C}}\widehat{\sigma}_c^2(t)\left(\sum_{i\in s_m^{(c)}(t)}\frac{1-\pi_i}{\pi_i^2}\right), \tag{5.66}$$

avec $\tilde{\mathrm{Y}}_i(t) = \mathrm{Y}_i(t)$ pour $i \in s_r(t)$ et $\tilde{\mathrm{Y}}_i(t) = \mathrm{Y}_i^*(t)$ pour $i \in s_m^{(c)}(t)$. La variance $\sigma_c^2(t)$ est remplacée par un estimateur obtenu en considérant la moyenne des carrés des différences, pour les données complètement observées, entre les valeurs interpolées et les valeurs observées. L'expression de la variance d'imputation se simplifie alors

$$\begin{aligned}\mathbb{V}_m\left(\widehat{t}_{int}(t) - \widehat{t}_{\mathrm{Y}}(t)|s,s_r\right) &= \sum_{i\in s_m(t)}\mathbb{V}_m\left(\sum_{\ell=1}^{\mathrm{L}}\varphi_{ii}(t,t_\ell)\frac{\mathrm{Y}_i(t_\ell)}{\pi_i} - \frac{\mathrm{Y}_i(t)}{\pi_i}|s,s_r\right)\\ &= \sum_{c=1}^{\mathrm{C}}\sum_{i\in s_m^{(c)}(t)}\frac{1}{\pi_i^2}\sum_{\ell,\ell'=1}^{p}\sigma_c^2(t_\ell,t'_\ell)\left(\varphi_{ii}(t,t_\ell) - \mathbb{1}_{\{t=t_\ell\}}\right)\left(\varphi_{ii}(t,t_{\ell'}) - \mathbb{1}_{\{t=t_{\ell'}\}}\right).\end{aligned}$$

En supposant qu'il n'y ait pas de dépendance temporelle dans les résidus du modèle de superpopulation, *i.e.* $(\mathbb{C}\mathrm{ov}_m(\mathrm{Y}_i(t),\mathrm{Y}_i(v)) = \sigma_c^2(t)\mathbb{1}_{\{t=v\}})$, les expressions précédentes deviennent beaucoup plus simples et un estimateur est obtenu en remplaçant $\sigma_c^2(t)$ par l'estimateur $\widehat{\sigma}_c^2(t)$ décrit précédemment,

$$\widehat{\mathbb{V}}_{\text{IMP}}\left(\widehat{t}_{int}(t) - \widehat{t}_{\mathrm{Y}}(t)\right) = \sum_{c=1}^{\mathrm{C}}\sum_{i\in s_m^{(c)}(t)}\frac{1}{\pi_i^2}\sum_{\ell=1}^{\mathrm{L}}\widehat{\sigma}_c^2(t_\ell)\left(\varphi_{ii}(t,t_\ell) - \mathbb{1}_{\{t=t_\ell\}}\right)^2. \tag{5.67}$$

En utilisant (5.57), un estimateur de la variance mixte est obtenu par

$$\widehat{\mathbb{V}}_{\text{MIX}}(t_\ell) = 2\sum_{c=1}^{\mathrm{C}}\sum_{i\in s_m^{(c)}(t)}\frac{1-\pi_i}{\pi_i}\,\widetilde{w}_{i,\ell}(t_\ell)\widehat{\sigma}_c^2(t_\ell), \tag{5.68}$$

avec $\widetilde{w}_{i,\ell}(t_\ell)$ définie dans (5.56). L'estimateur total de variance est obtenu en additionnant les expressions de (5.66), (5.67) et (5.68).

Estimateur de variance pour l'imputation par l'estimateur à noyau

On s'intéresse maintenant à l'imputation par l'estimateur lissé à noyau. On considère le modèle de superpopulation simplifié et on suppose que, pour $i \in \mathrm{U}_c$,

$$\mathbb{E}_m(\mathrm{Y}_i(t)) = \mu_c(t)$$
$$\mathbb{C}\mathrm{ov}_m(\mathrm{Y}_i(t), \mathrm{Y}_j(v)) = \sigma_c^2(t)\mathbb{1}_{\{t=v\}}\mathbb{1}_{\{i=j\}},$$

avec la moyenne conditionnelle $\mu_c(t)$ et la variance conditionnelle $\sigma_c^2(t)$ qui ne dépendent pas de l'unité i mais peuvent varier d'une classe d'imputation à l'autre. On suppose de plus que la vraie courbe $\mu_c(t)$ est une fonction lisse et donc que le biais peut être considéré comme négligeable par rapport à la variance si la fenêtre est suffisamment petite et la grille de discrétisation est suffisamment dense (voir les hypothèses faites en annexe).

Comme précédemment, la variance d'échantillonnage peut être estimée de la manière suivante,

$$\widehat{\mathbb{V}}_{\mathrm{SAM}}(t) = \sum_{i,j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \tilde{\mathrm{Y}}_i(t)\tilde{\mathrm{Y}}_j(t) + \sum_{c=1}^{\mathrm{C}} \widehat{\sigma}_c^2(t) \left(\sum_{i\in s_m^{(c)}(t)} \frac{1-\pi_i}{\pi_i^2} \right), \tag{5.69}$$

avec $\tilde{\mathrm{Y}}_i(t) = \mathrm{Y}_i(t)$ si $i \in s_r(t)$ et $\tilde{\mathrm{Y}}_i(t) = \widehat{\mu}_{c,\mathrm{Ha}}^{(1)}(t)$ si $i \in s_m(t) \cap \mathrm{U}_c$ et $\widehat{\sigma}_c^2(t)$ la moyenne des carrés des différences entre les valeurs observées à un instant t pour la classe U_c et $\widehat{t}_{\mathrm{K}_c}(t)/\mathrm{N}_c$.

En supposant que le biais est négligeable, on propose comme dans (5.52), l'approximation suivante pour le surcroît de variance dû à l'imputation,

$$\mathbb{E}_p\mathbb{E}_{\mathrm{R}}\mathbb{E}_m\left(\widehat{t}_{\mathrm{K}}(t) - \widehat{t}_{\mathrm{Y}}(t)|s, s_r\right)^2 \approx \mathbb{E}_p\mathbb{E}_{\mathrm{R}}\mathbb{V}_m\left(\widehat{t}_{\mathrm{K}}(t) - \widehat{t}_{\mathrm{Y}}(t)|s, s_r\right),$$

et, avec (5.55), on peut considérer l'estimateur suivant pour $\mathbb{V}_{\mathrm{IMP}}(t)$,

$$\widehat{\mathbb{V}}_{\mathrm{IMP}}(t) = \widehat{\mathbb{V}}_m\left(\widehat{t}_{\mathrm{K}}(t) - \widehat{t}_{\mathrm{Y}}(t)|s, s_r\right) = \sum_{c=1}^{\mathrm{C}} \sum_{\ell=1}^{\mathrm{L}} \widehat{\sigma}_c^2(t_\ell) \left(\sum_{i\in s^{(c)}} \widetilde{w}_{i,\ell}^2(t) \right). \tag{5.70}$$

Ici encore, une estimation de la variance mixte peut être obtenue en remplaçant la variance dans le modèle de superpopulation par une estimation,

$$\widehat{\mathbb{V}}_{\mathrm{MIX}}(t_\ell) = 2\sum_{c=1}^{\mathrm{C}} \sum_{i\in s_m^{(c)}(t)} \frac{1-\pi_i}{\pi_i} \widetilde{w}_{i,\ell}(t_\ell)\widehat{\sigma}_c^2(t_\ell). \tag{5.71}$$

L'estimateur final de variance pour l'estimateur par imputation du noyau est obtenu en additionnant (5.69), (5.70) et (5.71).

5.4 Application aux données réelles de consommation électriques

5.4.1 Données, plans de sondage et scénarios non réponse

Nous avons travaillé sur les données irlandaises déjà présentées dans la section 3.6.1 du chapitre 3. La population d'intérêt (3994 ménages), la période d'étude (18 au

24 janvier 2010) et les variables explicatives sont les mêmes que précédemment. Pour que nos courbes soient plus lisses et moins bruitées, nous avons réalisé un lissage par moyenne mobile d'ordre 9 sur chacune des courbes d'origine.

Protocole de simulation

Afin d'évaluer la qualité de nos estimateurs de la courbe totale et de la comparer à celles de méthodes d'imputation plus classiques, fréquemment utilisées à EDF, on considère des plans de sondage et des scénarios de non réponse variés (avec des séquences manquantes longues ou courtes, survenant simultanément ou non). Pour des raisons de confidentialité, on ne peut pas donner plus de détails sur la structure réelle de la non réponse ni même sur son taux. Cependant les scénarios de non réponse proposés ont été conçus pour couvrir l'ensemble des cas rencontrés dans la réalité et pour pouvoir évaluer en particulier l'impact de la longueur des séquences manquantes sur les performances respectives des différentes méthodes.

Plus précisément, deux plans de sondages sont considérés :

— SAS : sondage aléatoire simple

— STR SJ0 : Sondage stratifié, avec 5 strates construites en fonction de la consommation moyenne de chaque client sur le semestre précédent (Juillet-Décembre 2009) et sondage aléatoire simple dans chaque strate.

Pour chaque plan de sondage, on sélectionne des échantillons de taille $n = 400$. Pour le plan stratifié, on définit des limites de strates de façon à ce que le total de la variable de stratification soit le même dans chaque strate pour la population. Les échantillons sont sélectionnés par allocation proportionnelle à la taille dans chaque strate.

On considère différents scénarios de non réponse.

- Non réponse totale : 10% des unités, sélectionnées par un processus de sélection MCAR (Missing Completely At Random), ont leur courbe intégralement manquante.
- Non réponse d'une journée : les couples (journée, unité) affectés par la non réponse sont sélectionnés aléatoirement avec une probabilité uniforme de 10% MCAR
- Petits trous : on sélectionne aléatoirement les couples (instant de début de la séquence, unité) avec une probabilité uniforme de 10% de valeurs manquantes. Les longueurs de ces "petits trous" se répartissent de la manière suivante : 50 % des instants manquants correspondent à des trous de longueur 1, et 10 % des instants correspondent respectivement aux longueurs 2, 3, 4, 5 et 6.
- Non réponse simultanée : pour 50 instants sélectionnés aléatoirement, 67,2% des unités sont manquantes. Ces unités sont sélectionnées aléatoirement et ne sont pas forcément identiques d'un instant à l'autre. On a par construction $\frac{50}{336} \times 0.672 = 10\%$ de valeurs manquantes, MCAR.
- Scénario mixte avec 10% (resp. 20 %) : combinaison de l'ensemble des scénarios précédents, avec 5% (resp. 10%) de non réponse totale, 1% (resp. 2%) de non réponse d'une journée, 3% (resp. 6%) de petits trous, et 1% (resp. 2%) de non réponse simultanée. Cela fait donc un total d'*environ* (voir remarque plus bas) 10% (resp. 20%) de non réponse MCAR.

Remarque 11. *Pour les petits trous ainsi que le scénario mixte, les trous de différentes longueurs sont simulés indépendamment et il est donc possible que deux trous de longueurs différentes se "chevauchent". Le taux de non réponse finalement obtenu n'est donc pas exactement de 10% mais d'un peu moins : 9,61% en moyenne pour le scénario mixte à 10%, 18.6% pour le scénario mixte à 20%, et 9,63% pour le scénario avec 10% de petits trous. Néanmoins, on considère que ça n'est pas grave puisqu'on cherche en priorité à comparer les méthodes entre elles plutôt que les scénarios entre eux.*

Pour alléger le protocole de test, on ne combine pas les deux plans de sondage avec l'ensemble des scénarios de non réponse, mais on se limite à 7 scénarios ci-dessous :

- Sondage aléatoire simple, scénario mixte, 10% de valeurs manquantes (SAS - Mixte).
- Sondage stratifié, scénario mixte, 10% de valeurs manquantes (Strat - Mixte).
- Sondage stratifié, non réponse d'une journée, 10% de valeurs manquantes (Strat - Jour).
- Sondage stratifié, non réponse totale, 10% de valeurs manquantes (Strat - Total).
- Sondage stratifié, petits trous, 10% de valeurs manquantes (Strat - Petits).
- Sondage stratifié, non réponse simultanée, 10% de valeurs manquantes (Strat - Synchrone)
- Sondage stratifié, scénario mixte, 20% de valeurs manquantes (Strat - Mixte - 20)

5.4.2 Estimateurs comparés et indicateurs de performance

On compare les estimateurs suggérées ici avec d'autres méthodes fréquemment utilisées en pratique :

- Moyenne de classe : chaque valeur manquante est remplacée par la moyenne des valeurs observées dans la classe d'imputation (pour l'instant considéré) (noté *impute moyenne classe*).
- Interpolation linéaire des petits trous (jusqu'à trois heures, c'est-à-dire 6 points consécutifs exclus) et moyenne de classe sinon. (*interpo ou moy*).

Les nouveaux estimateurs considérés sont les suivants :

- Lissage à noyau présenté dans l'équation (5.12) , notée *estim noyaux* . Le choix de la fenêtre η est réalisé en minimisant le critère de validation croisée défini par l'équation (5.20).
- Imputation par lissage à noyau présentée dans (5.40), notée *impute noyaux*. Le choix de la fenêtre η est réalisé en minimisant le critère de validation croisée présenté dans (5.20).
- Interpolation linéaire des différences à la moyenne présentée dans (5.43), notée *interpo diff moy*.
- Imputation des plus proches voisins, avec $k = 1, 2$ ou 5 voisins (notée *ppv*), décrite dans ((5.41)).

On utilise C = 5 classes d'imputation, qui correspondent aux strates du sondage stratifié. Comme référence, on considère également l'estimateur de Horvitz-Thompson en l'absence de non réponse (noté *complet*).

Afin de quantifier les performances de nos méthodes d'estimation de courbe moyenne, on tire un grand nombre E = 1000 d'échantillons dans la population et, pour chacun de ces échantillons, on simule de la non réponse selon les différents scénarios présentés précédemment. On évalue ensuite les performances des différents estimateurs à l'aide des indicateurs déjà introduits dans les chapitres précédents : le Biais Relatif (RB) et l'erreur quadratique moyenne (MSE) définis pour un instant $t_l, \quad l = 1, \ldots, \mathrm{L}$ dans les équations (3.77) ainsi que l'Efficacité Relative (RE) définie dans l'équation (3.79) avec pour estimateur de référence $\hat{\mu}_{\mathrm{Y}}^{0}$ l'estimateur de Horvitz-Thompson lorsqu'il n'y a pas de données manquantes (noté *complet* dans les Tables et les Figures). Cet indicateur permet de faciliter les comparaisons : un RE de 100 correspond à une précision égale à celle que l'on aurait eue sans non réponse et plus RE est bas plus cela signifie que l'estimateur est précis.

Comme précédemment, on synthétise finalement chacun de ces indicateurs en prenant sa moyenne sur l'ensemble des instants de discrétisation. On s'intéressera aussi aux temps de calcul des différents estimateurs.

Afin de mesurer les performances de nos estimateurs de variance, on procède de la même manière. On tire un grand nombre E = 10000 d'échantillons de taille $n = 400$ dans la population et, pour chacun de ces échantillons, on simule de la non réponse selon les différents scénarios présentés précédemment. On compare ensuite la variance estimée avec l'approximation Monte Carlo de la variance déduite des simulations par :

$$\mathbb{V}_{\mathrm{MC}}(\hat{\mu}_{\mathrm{Y}}(t_\ell)) = \frac{1}{\mathrm{E}-1}\sum_{e=1}^{\mathrm{E}}(\hat{\mu}_{\mathrm{Y}}^{e}(t_\ell)) - \mathbb{E}_{\mathrm{MC}}[\hat{\mu}_{\mathrm{Y}}(t_\ell)])^2$$

Plus précisément, notre indicateur est la moyenne sur l'ensemble des instants de l'écart relatif entre cette quantité et la valeur de l'estimateur,

$$\mathrm{RB}(\hat{\mathbb{V}}(\hat{\mu}_{\mathrm{Y}}))) = 100\frac{1}{\mathrm{L}}\sum_{\ell=1}^{\mathrm{L}}\left(\frac{\hat{\mathbb{V}}(\hat{\mu}_{\mathrm{Y}}(t_\ell)) - \mathbb{V}_{\mathrm{MC}}(\mu_{\mathrm{Y}}(t_\ell))}{\mathbb{V}_{\mathrm{MC}}(\mu_{\mathrm{Y}}(t_\ell))}\right). \tag{5.72}$$

Si nos estimateurs sont effectivement sans biais, on s'attend à ce que ce critère soit faible en valeur absolue. On réalise les tests uniquement sur le sondage aléatoire stratifié et le scénario mixte (Strat-Mixte), pour $n = 400$.

5.4.3 Aspects pratiques de l'implémentation des estimateurs de courbe moyenne ou totale ainsi que des estimateurs de variance associés

Dans l'implémentation pratique, nous avons apporté certaines modifications par rapport aux méthodes d'imputation présentées plus haut. Ainsi, l'interpolation linéaire ainsi que la méthode basée sur les plus proches voisins ne peuvent parfois pas être mises en œuvre, par exemple lorsque les premiers ou les derniers instants de la période sont manquants. Dans ce cas, on impute les valeurs manquantes concernées par

la moyenne de classe. En outre, pour l'ensemble des méthodes d'imputation, lorsque l'ensemble des unités de la classe sont manquantes, on réalise l'imputation en utilisant l'ensemble de l'échantillon. Enfin, pour la méthode des plus proches voisins, lorsqu'il y a au moins un voisin dans la classe d'imputation mais moins que le nombre k souhaité, on impute par la moyenne des voisins présents.

Pour l'estimation des variances de nos estimateurs, des ajustements ont également été réalisés. En particulier, pour l'ensemble des méthodes, il est nécessaire d'estimer la variance conditionnelle de chaque unité $\hat{\sigma}_i^2$, qui dépend de la classe d'imputation. Or, les unités répondantes ne peuvent pas être affectées à une classe d'imputation, car celles-ci sont construites en fonction de la longueur des séquences manquantes, croisée avec les autres informations auxiliaires. Pour ces unités, nous avons donc choisi d'estimer les variances conditionnelles par la moyenne pondérée des variances de chacune des classes d'imputation correspondant aux différentes longueurs de séquences manquantes (pour l'information auxiliaire des unités concernées), les poids étant proportionnels aux fréquences empiriques de l'occurrence de ces différentes longueurs de séquences.

Par ailleurs, pour la méthode d'imputation des plus proches voisins, l'estimation des variances conditionnelles des répondants σ_i^2 aurait nécessité de trouver les plus proches voisins de ces unités, pour l'ensemble des longueurs de séquences manquantes possibles, ce qui aurait rallongé considérablement les temps de calcul. Afin d'être plus rapide, on estime ces variances conditionnelles par la moyenne des variances conditionnelles estimées pour les non répondants de la classe d'imputation : il s'agit alors implicitement d'une moyenne des variances conditionnelles des différentes longueurs de séquences manquantes, pondérées par leur fréquence d'occurrence empirique. L'estimation de variance n'augmente alors presque pas les temps de calcul par rapport à l'estimation simple de courbes moyennes.

Enfin, comme mentionné plus haut, dans les cas où l'imputation par plus proche voisin ou interpolation linéaire n'est pas réalisable, on impute les valeurs manquantes par moyennes de classe. Dans le calcul de variance, cette adaptation est prise en compte en considérant dans les termes de variance mixte et de variance d'imputation des termes correspondant à une imputation par les plus proches voisins, pour un nombre k de voisins égal au nombre de répondants de la classe. Cela correspond en effet exactement à l'imputation par la moyenne de classe.

Par ailleurs, pour la construction de l'estimateur par lissage à noyau (utilisé directement ou pour imputer les valeurs manquantes), le choix de la fenêtre est capital. Il est réalisé selon le critère de validation croisée (5.20) qui permet de tenir compte à la fois du plan de sondage et du processus de non réponse, et d'optimiser la qualité de l'estimation de la courbe moyenne et non de chaque courbe individuelle. La recherche de ce paramètre est réalisée par grid-search, pour des valeurs comprises entre 1 et 4 instants par pas de 0.5.

5.4.4 Résultats

Les résultats des comparaisons des différents estimateurs de courbes totales sont résumés dans les Tables 5.1 à 5.6 et dans les Figures 5.3 à 5.9. Les principales conclusions sont résumées ici.

Tout d'abord, comme le montrent les Tables 5.1 et 5.2, dans nos simulations le biais relatif est plus faible que 1%, pour l'ensemble des stratégies d'imputation et des scénarios de non réponse à l'exception de la non réponse simultanée. Pour ce dernier scénario, on constate en effet des sous-estimations systématiques entre 1 et 2 % pour la méthode du plus proche voisin. Il est intéressant de remarquer que le biais relatif augmente avec le nombre de voisins. L'analyse des courbes des biais relatifs sur la Figure 5.9 montre que les sous-estimations les plus importantes ont lieu entre 4h et 7h du matin tous les jours de la semaine. La faiblesse des biais dans l'ensemble des scénarios est probablement liée au fait que dans l'ensemble des scénarios testés, les valeurs sont *Missing Completely at Random*.

Par ailleurs, en observant les MSE des Table 5.3 et 5.4, on remarque, comme attendu, que les erreurs d'estimation sont beaucoup plus importantes pour le Sondage Aléatoire Simple que pour le sondage stratifié. Pour le sondage stratifié, les erreurs sont plus importantes pour la non réponse totale ainsi que la non réponse d'un jour, c'est-à-dire en présence de séquences longues de valeurs manquantes. A contrario, les erreurs d'estimation sont les plus faibles pour les plus petites séquences de valeurs manquantes.

En examinant les efficacités relatives résumées dans les Tables 5.5 et 5.6, on constate que le classement relatif des différentes méthodes dépend des scénarios de valeurs manquantes. On remarque tout d'abord que l'imputation par la moyenne de classe est presque toujours la plus mauvaise méthode, à part pour le scénario de non réponse totale, pour lequel la seule technique alternative possible est l'estimateur à noyau. L'interpolation linéaire et les plus proches voisins donnent en général de bons résultats et sont fréquemment les meilleures méthodes pour les séquences courtes ainsi que le scénario mixte. On remarque cependant que, dans le scénario d'occurrence simultanée, l'interpolation linéaire est largement meilleure que les plus proches voisins. Dans ce cas, la performance décroît lorsque le nombre de voisins augmente (de $k = 1$ à $k = 5$). Pour des séquences manquantes d'une journée entière, le plus proche voisin est la méthode la plus performante et il semble préférable de considérer plus d'un voisin.

En ce qui concerne les temps de calcul, comme le montre la Table 5.8, la moyenne de classe et les interpolations linéaires sont extrêmement rapides à mettre en œuvre (moins de trois dixièmes de seconde, voire d'un dixième de seconde pour la moyenne de classe) tandis que les méthodes des plus proches voisins et des estimateurs à noyau et imputation par l'estimateur à noyau sont respectivement 40 et 160 fois plus longues. L'estimation de variance ne ralentit que très peu l'implémentation de la méthode du plus proche voisin (10 secondes) ainsi que de l'imputation à la moyenne (0.4 secondes)). En revanche, elle fait passer le temps de l'estimation par interpolation linéaire ou moyenne de classe à 10 secondes.

Enfin, comme on peut le voir dans la Table 5.7, les estimateurs de MSE surestiment légèrement celui-ci (6% au minimum et 8.4% pour le plus proche voisin). Ces estimateurs sont donc plutôt conservatifs, ce qui n'est pas très gênant.

méthode	Strat - Mixte	SAS - Mixte	Strat - Petits	Strat - Total
complet	0.18	-0.04	-0.02	-0.01
ppv5	-0.29	-0.36	-1.07	-0.00
ppv2	-0.10	-0.22	-0.64	-0.00
ppv1	-0.03	-0.15	-0.47	-0.00
interpo ou moy	0.18	-0.01	0.02	-0.00
interpo diff moy	0.17	-0.02	-0.02	-0.00
estim noyaux	0.40	0.19	0.21	0.16
impute noyaux	0.20	-0.01	0.01	0.02
impute moyenne classe	0.18	-0.03	-0.01	-0.00

TABLEAU 5.1 – RB (%) pour chaque méthode d'estimation et chaque scénario (partie 1).

méthode	Strat - Jour	Strat - Synchrones	Strat Mixte 20 %
complet	0.18	-0.27	-0.24
ppv5	-0.19	-1.93	-1.27
ppv2	-0.03	-1.31	-0.98
ppv1	0.02	-1.02	-0.84
interpo ou moy	0.22	-0.25	-0.31
interpo diff moy	0.22	-0.24	-0.34
estim noyaux	0.45	-0.02	-0.09
impute noyaux	0.24	-0.23	-0.29
impute moyenne classe	0.22	-0.22	-0.33

TABLEAU 5.2 – RB (%) pour chaque méthode d'estimation et chaque scénario (partie 2).

méthode	Strat - Mixte	SAS - Mixte	Strat - Petits	Strat - Total
complet	3569	4962	3539	3678
ppv5	3783	5282	3560	4138
ppv2	3794	5303	3543	4138
ppv1	3815	5327	3550	4138
interpo ou moy	3815	5336	3542	4138
interpo diff moy	3825	5344	3562	4138
estim noyaux	3918	5364	3729	4249
impute noyaux	3961	5529	3908	4130
impute moyenne classe	3995	5567	3969	4138

TABLEAU 5.3 – MSE pour chaque méthode d'estimation et chaque scénario (partie 1).

méthode	Strat - Jour	Strat - Synchrones	Strat Mixte 20 %
complet	3595	3675	3681
ppv5	3925	4858	4335
ppv2	3992	4149	4283
ppv1	4071	3952	4289
interpo ou moy	4066	3678	4240
interpo diff moy	4066	4252	4280
estim noyaux	4167	3784	4667
impute noyaux	4056	3908	4551
impute moyenne classe	4066	4871	4669

TABLEAU 5.4 – MSE pour chaque méthode d'estimation et chaque scénario (partie 2).

méthode	Strat - Mixte	SAS - Mixte	Strat - Petits	Strat - Total
complet	100.00	100.00	100.00	100.00
ppv5	106.00	106.45	100.59	112.50
ppv2	106.33	106.87	100.12	112.50
ppv1	106.90	107.34	100.32	112.50
interpo ou moy	106.91	107.53	100.07	112.50
interpo diff moy	107.18	107.69	100.65	112.50
estim noyaux	109.80	108.10	105.36	115.52
impute noyaux	111.01	111.42	110.44	112.28
impute moyenne classe	111.96	112.18	112.16	112.50

TABLEAU 5.5 – RE (%) pour chaque méthode d'estimation et chaque scénario (partie 1).

méthode	Strat - Jour	Strat - Synchrones	Strat Mixte 20 %
complet	100.00	100.00	100.00
ppv5	109.17	132.19	117.78
ppv2	111.02	112.89	116.37
ppv1	113.24	107.53	116.52
interpo ou moy	113.08	100.07	115.19
interpo diff moy	113.08	115.68	116.29
estim noyaux	115.90	102.95	126.80
impute noyaux	112.81	106.34	123.64
impute moyenne classe	113.08	132.54	126.86

TABLEAU 5.6 – RE (%) pour chaque méthode d'estimation et chaque scénario (partie 2).

méthode	RBV$\hat{ar}$
complet	6.25
interpo ou moy	6.61
ppv1	8.38

TABLEAU 5.7 – Biais relatifs (en %) des estimateurs de variance, pour le sondage stratifié et le scénario mixte.

	méthode	sans variance	avec variance
1	interpo diff moy	0.26	0.27
2	interpo ou moy	0.25	9.73
3	impute moyenne classe	0.10	0.35
4	impute noyaux	36.56	
5	ppv1	9.44	10.30

TABLEAU 5.8 – Temps de calcul (en secondes) pour les différentes méthodes d'estimation, avec et sans calcul de variance.

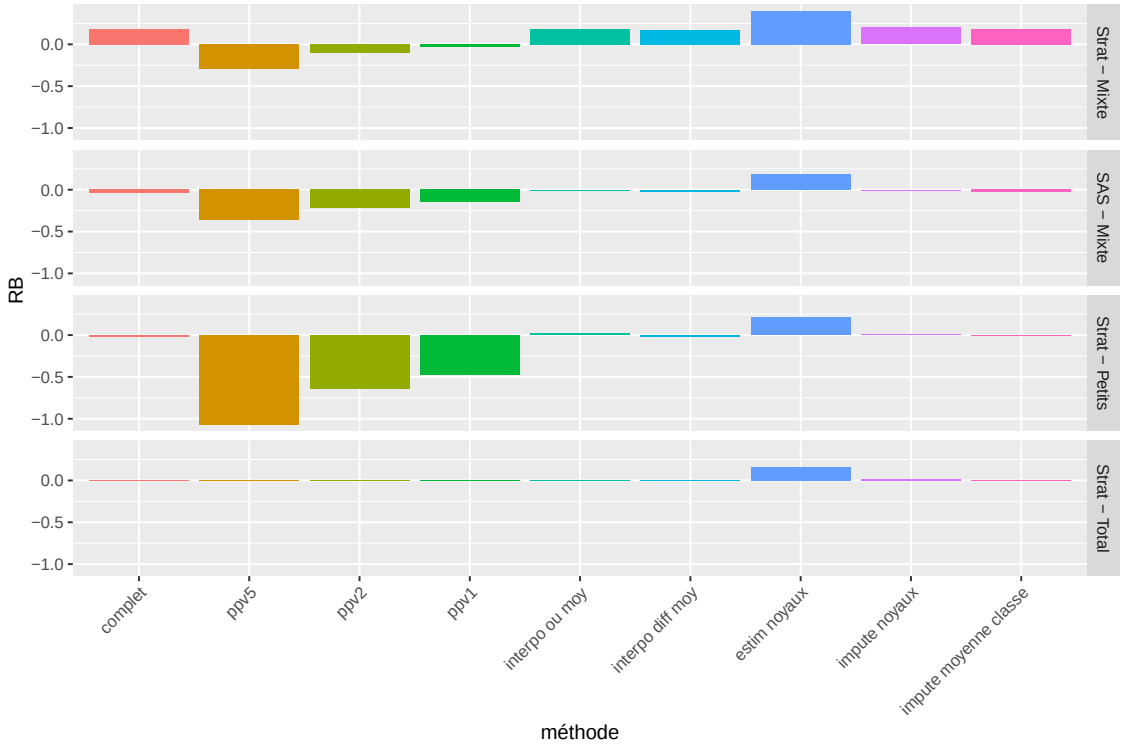

FIGURE 5.3 – Biais relatifs en % des différents estimateurs ponctuels, pour chacun des scénarios de test (partie 1).

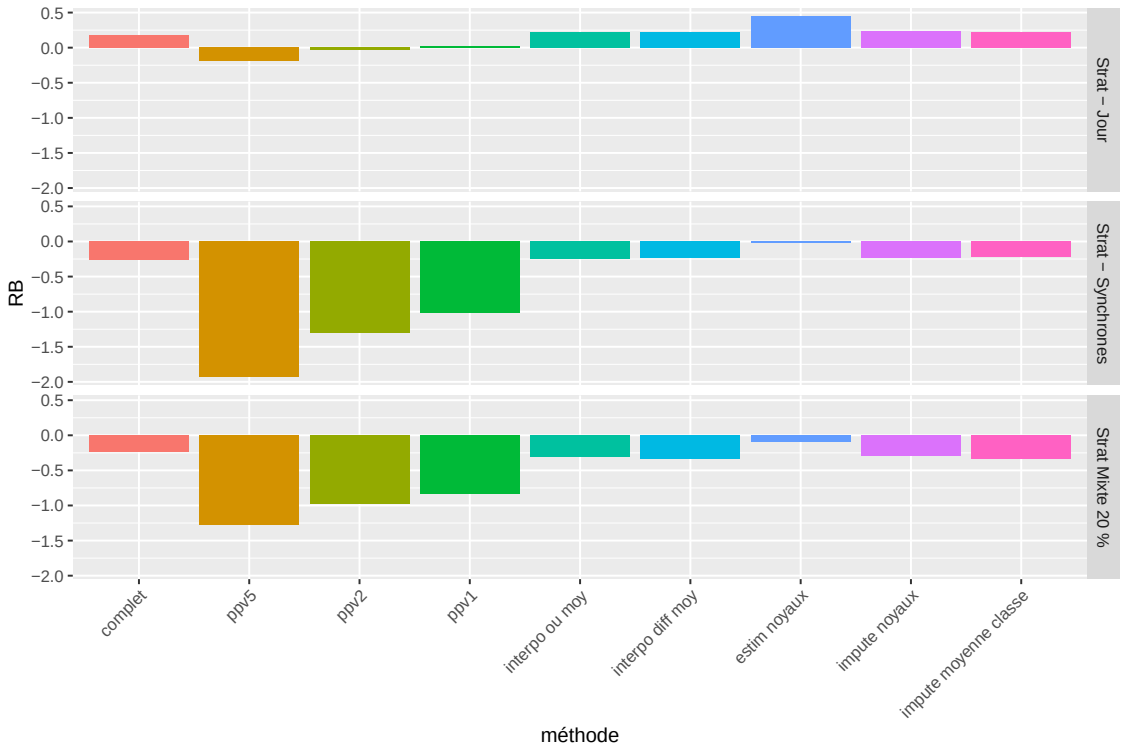

FIGURE 5.4 – Biais relatifs en % des différents estimateurs ponctuels, pour chacun des scénarios de test (partie 2).

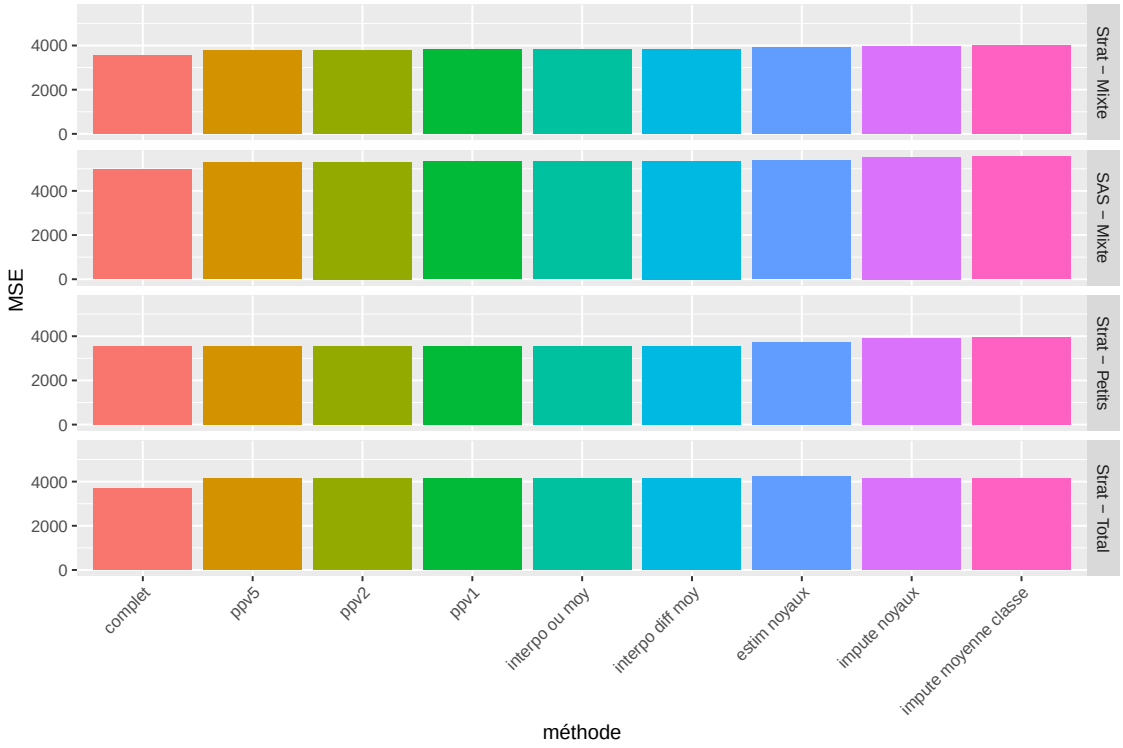

FIGURE 5.5 – Erreurs quadratiques moyennes (MSE) des différents estimateurs pour chacun des scénarios de test (partie 1).

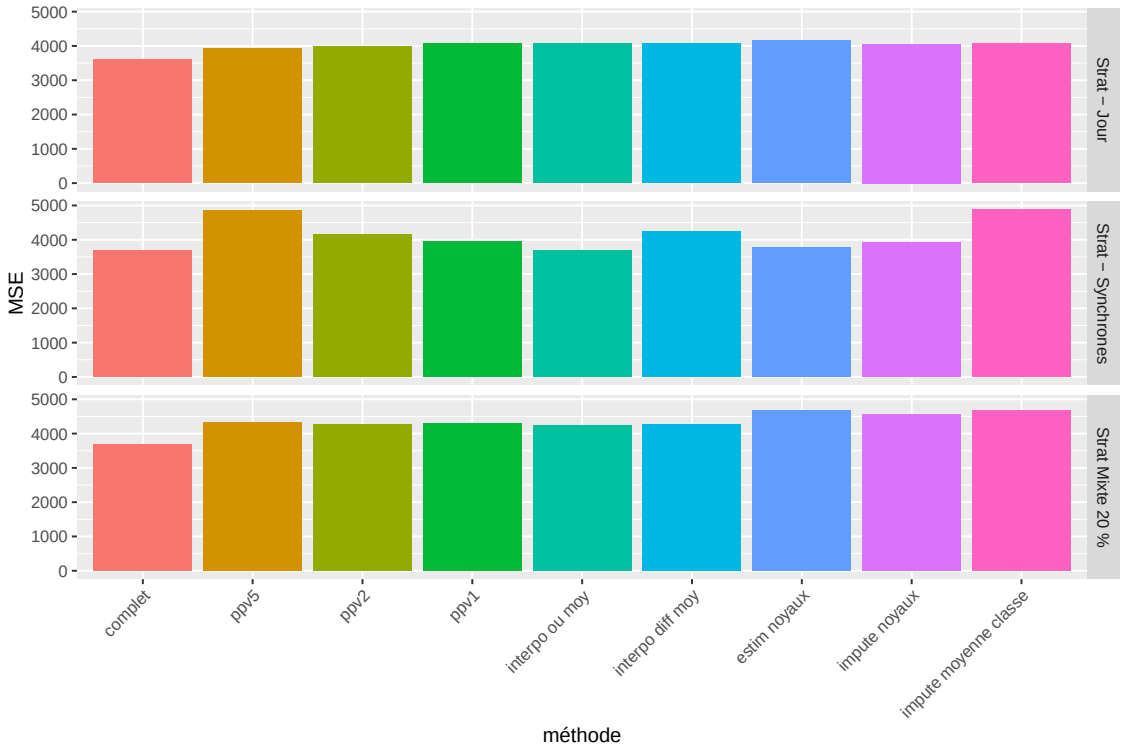

FIGURE 5.6 – Erreurs quadratiques moyennes (MSE) des différents estimateurs pour chacun des scénarios de test (partie 2).

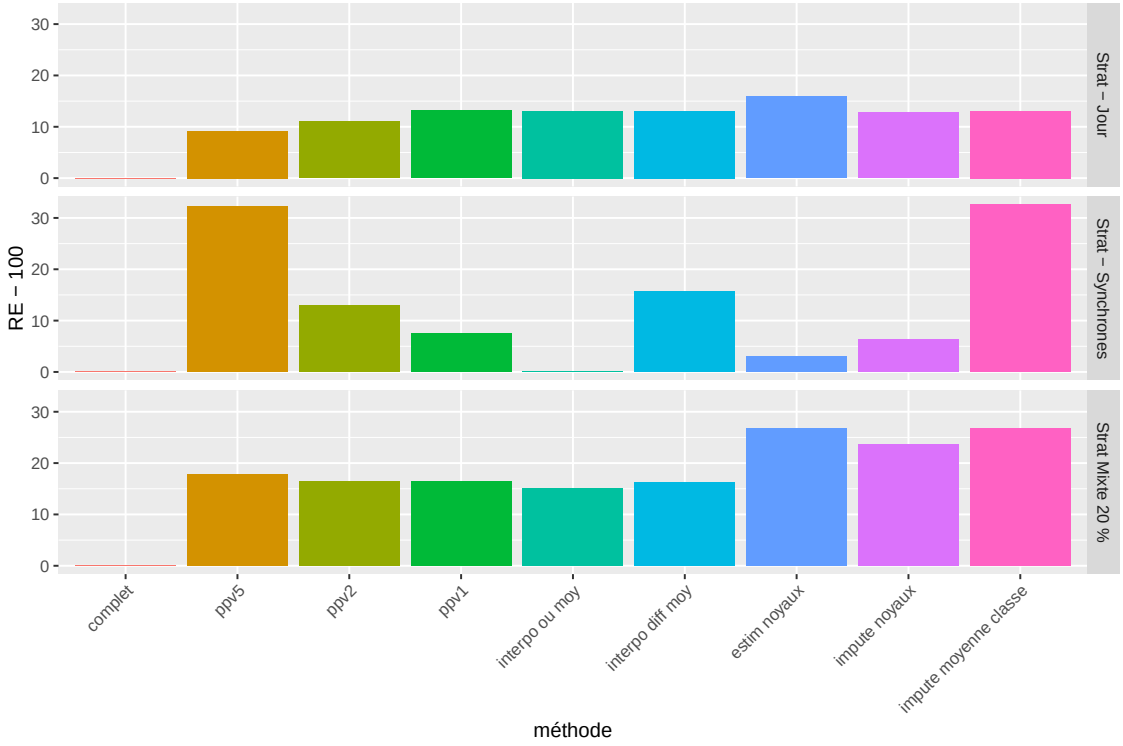

FIGURE 5.7 – Pertes d'efficacité relative en % (RE -1) des différents estimateurs, pour chacun des scénarios de test (partie 1).

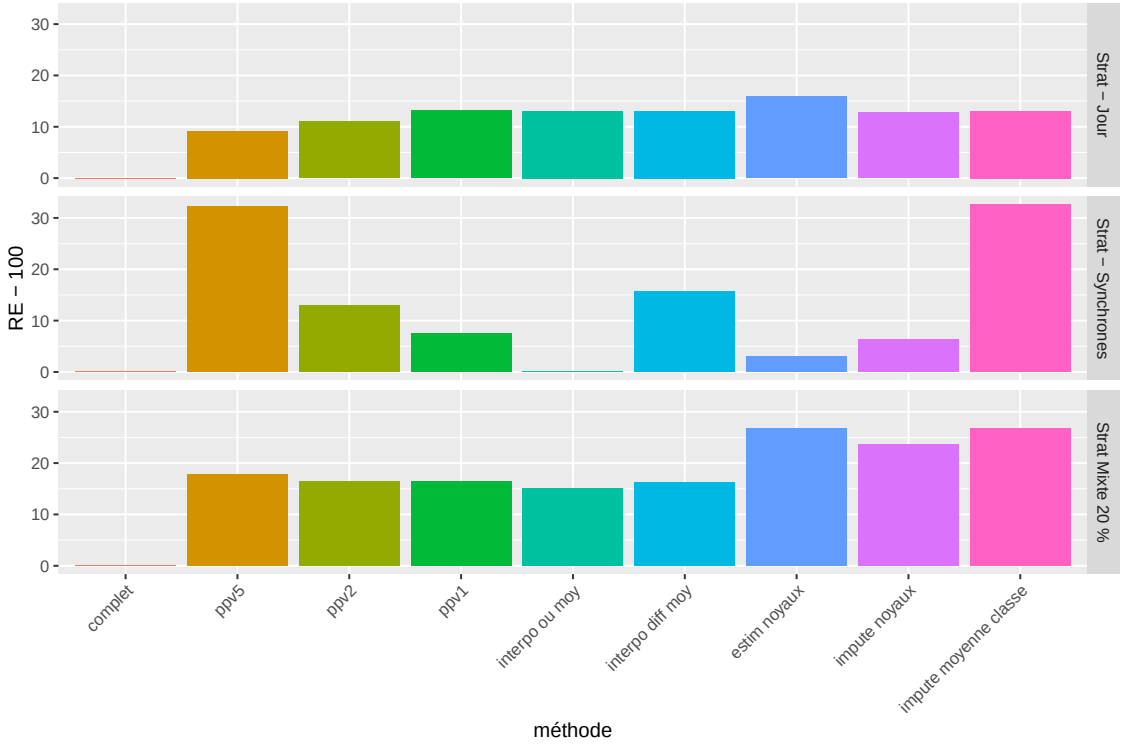

FIGURE 5.8 – Pertes d'efficacité relative en % (RE -1) des différents estimateurs, pour chacun des scénarios de test (partie 2).

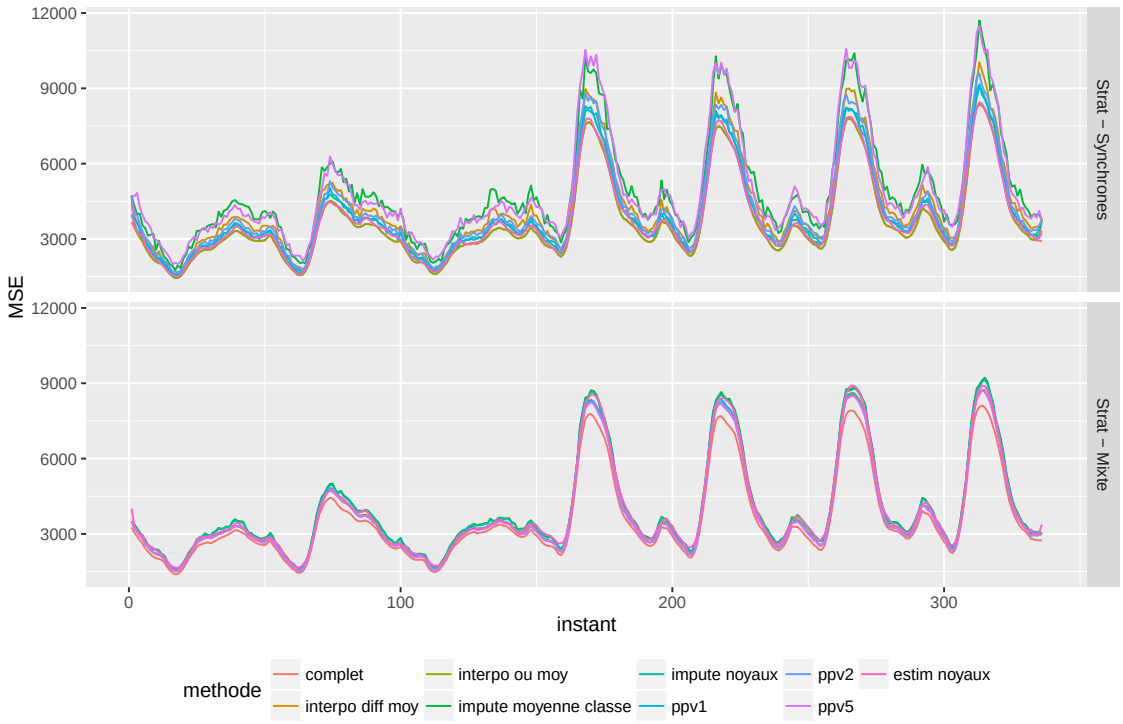

FIGURE 5.9 – Évolution du MSE au cours du temps pour les différents estimateurs, sondage stratifié et scénario mixte ou scénario de bugs simultanés.

5.5 Conclusions sur l'estimation de courbes en présence de valeurs manquantes

5.5.1 Conclusions méthodologiques

Dans ce chapitre, nous nous sommes intéressés à la question de l'estimation de courbe moyenne à partir d'un panel dont certaines courbes sont partiellement ou totalement inobservées. Pour répondre à cette problématique, similaire à celle de la non réponse partielle en sondages, nous avons proposé des méthodes d'estimation en présence de valeurs manquantes construites en étendant des techniques usuelles de sondages (telles que les plus proches voisins) au cadre des données fonctionnelles ou inversement en adaptant des techniques d'analyse des données fonctionnelles au contexte des sondages et de la non réponse. Notre première méthode consiste à approximer la courbe moyenne par lissage à noyau adapté au contexte des sondages et des valeurs manquantes. Le principe de la seconde méthode est d'utiliser cet estimateur à noyau pour approximer les courbes moyennes de chaque classe d'imputation puis à imputer les valeurs manquantes par ces approximations de moyennes de classe pour les instants considérés.

Une autre méthode d'imputation consiste à adapter la technique usuelle des plus proches voisins au contexte fonctionnel : pour cela on sélectionne un ensemble de voisins différents pour chaque séquence de valeurs manquantes, en fonction des valeurs de la courbe juste avant et juste après la séquence non observée. Les voisins peuvent donc changer au cours du temps afin de s'adapter au caractère très irrégulier des courbes de consommation électrique. Cependant, afin d'éviter que l'imputation ne repose que sur des similarités de très court terme entre les courbes, on recommande de construire au préalable des classes d'imputation en se servant de caractéristiques plus globales des unités, afin de prendre en compte également des critères plus globaux de ressemblances entre unités. Comme l'ensemble de chaque séquence manquante est imputée conjointement, on peut espérer que cette méthode permet de préserver les corrélations temporelles des courbes imputées, ainsi que de la courbe moyenne estimée.

Enfin, la quatrième méthode testée est l'interpolation de la différence à la courbe moyenne. Cette technique est très simple à mettre en œuvre et permet de tenir compte simultanément des tendances globales observées sur la population (par la courbe moyenne) et des spécificités de chaque individu déduites des portions observées des courbes (par les différences à cette moyenne).

Pour chacune de ces méthodes, nous avons proposé une approximation de variance. Pour cela, pour l'ensemble des méthodes d'imputation, nous avons développé une approche unifiée en nous inspirant de l'imputation composite décrite dans Beaumont and Bissonnette (2011).

L'ensemble des méthodes proposées dans ce chapitre ont été mises en œuvre et comparées sur le jeu de données des courbes de consommation irlandaises déjà utilisé dans le chapitre 3, pour différents plans de sondage et différents scénarios de valeurs manquantes. La principale conclusion de ces tests est que la méthode d'imputation à mettre en œuvre dépend de la longueur de la portion de courbe inobservée à compléter : en effet, l'interpolation linéaire simple (sans différence à la moyenne) est très

performante sur les séquences courtes et possède en outre l'avantage indéniable d'être extrêmement rapide à mettre en œuvre. Néanmoins à mesure que la longueur de la séquence considérée augmente, ses performances se dégradent et l'estimateur par plus proche voisin devient alors préférable. On choisira plutôt 1 ou 2 voisins, car les tests ont montré que pour 5 voisins les performances de la méthode pouvaient sensiblement se dégrader lorsque les séquences manquantes interviennent de manière simultanée. L'estimateur à noyau est souvent un peu meilleur que la moyenne de classe, cependant il est beaucoup plus lourd à mettre en œuvre et aux performances proches. Enfin, dans nos tests, les estimateurs de variance proposés semblent souffrir d'un léger biais de surestimation.

5.5.2 Perspectives

En pratique, il est souvent nécessaire de traiter conjointement le problème de la non réponse ainsi que d'autres problématiques, notamment la robustesse aux unités influentes et l'estimation pour des petits domaines qui sont abordés dans cette thèse. En effet, il est très fréquent que nos jeux de données contiennent des courbes partiellement ou totalement manquantes, or les méthodes présentées dans les chapitres précédents sont conçues pour des courbes intégralement observées et l'adaptation au cas des courbes incomplètes n'est pas forcément triviale. Un des principaux enseignements de ce chapitre est qu'il semble pertinent d'imputer les petites séquences de valeurs manquantes par interpolation linéaire.

La question de la méthode d'imputation à mettre en œuvre est cependant plus complexe pour les séquences plus longues. En effet, les chapitres précédents montrent que, dans notre contexte d'estimation de courbes moyennes, les méthodes les plus performantes sont en général celles qui prennent en compte l'aspect fonctionnel de la problématique, par exemple à l'aide d'une ACP robuste sur les instants de discrétisation. Or dans ce cadre un seul instant de discrétisation sans mesure induit des valeurs manquantes sur l'ensemble des composantes principales pour l'unité considérée. Il apparaît donc souhaitable d'imputer la courbe manquante par l'une des méthodes proposées dans ce chapitre. Néanmoins, dans le contexte de l'estimation sur des petits domaines l'imputation peut avoir des conséquences sur l'estimation des modèles, et en particulier des modèles linéaires mixtes, qu'il semble délicat de quantifier. Enfin, on remarquera que les méthodes d'imputation sont à utiliser pour la non réponse partielle et non pas la non réponse totale, or dans notre contexte fonctionnel la limite entre non réponse partielle et non réponse totale dépend de la fenêtre temporelle choisie. En effet, le fait de raisonner sur de plus petites fenêtres temporelles peut conduire à plus de non réponse totale et moins de non réponse partielle pour chaque période considérée.

D'un point de vue méthodologique, une piste de réflexion pourrait être par exemple la quantification de l'erreur quadratique moyenne pour l'estimation robuste en présence de valeurs manquantes. Une autre piste de réflexion serait l'intégration de la connaissance de résumés temporellement agrégés de la courbe de charge partiellement inobservée dans la procédure d'imputation. En effet, en plus des courbes de consommation dont certaines parties sont manquantes, on relève fréquemment

des index de consommation c'est-à-dire des cumuls de consommation depuis l'origine. Ces index sont cependant disponibles à une maille temporelle moins fine que la courbe de charge (journalière, mensuelle, voire semestrielle suivant les cas). Par simple soustraction on est donc à même de déduire les consommations sur la période de temps séparant des index et il semble pertinent d'exploiter cette information pour imputer les valeurs manquantes. Il s'agirait alors non plus d'estimer la valeur de la courbe aux instants manquants mais uniquement sa forme afin de "ventiler" les consommations manquantes déduites des index. On peut donc se demander comment réaliser au mieux cette estimation.

Enfin, nous avons cherché ici à proposer des méthodes d'estimation de courbes moyennes ou totales en présence de non réponse, or il pourrait également être intéressant de chercher à imputer les courbes de façon à estimer du mieux possible chaque valeur manquante. Plutôt qu'un problème de sondage, on serait alors confrontés à un problème d'estimation (sur des données issues d'un plan de sondage aléatoire dont il faudra toutefois tenir compte). En effet, il est fréquent que l'on souhaite disposer de telles méthodes d'imputation, par exemple lorsque l'on souhaite proposer à un client final un service impliquant la mise à disposition ou l'affichage de sa courbe de consommation électrique personnelle. Dans ce cas, l'affichage de valeurs manquantes est évidemment du plus mauvais effet. Dans ce contexte, l'enjeu est alors non seulement de produire des estimations précises mais surtout "crédibles", c'est-à-dire qui ressemblent visuellement à de "vraies" courbes individuelles, qui par nature sont en général beaucoup plus irrégulières que des courbes moyennes par exemple. La technique du plus proche voisin telle que présentée ici n'est pas non plus pertinente dans ce contexte car il existe des différences très fortes de niveau entre unités qui engendreraient des "sauts" dans la courbe imputée. Une piste potentielle pourrait donc être de découper la courbe en plus petites périodes, par exemple des jours, et à appliquer la méthode de recherche des plus proches voisins sur l'historique du client, en recherchant les jours qui ressemblent le plus à chaque jour partiellement inobservé.

Annexe A : détails techniques sur l'approximation de variance pour l'estimateur à noyau en présence de non réponse

A1. On suppose que la fonction μ est β-Hölder. Il existe $\beta \in]0,1]$ et une constante Λ tels que $\forall (t,u) \in [0,\mathrm{T}],\ |\mu_{\mathrm{Y}}(t) - \mu_{\mathrm{Y}}(u)| \leq \Lambda |t-u|^{\beta}$.

A2. On suppose que le noyau K est une fonction continue positive à support compact $[-1,1]$.

A3. On suppose que les instants de discrétisation $0 = t_1 < t_2 < \cdots < t_{\mathrm{L}} = \mathrm{T}$ sont équidistants, $t_l = (l-1)/(\mathrm{L}-1),\ l = 1,\ldots,\mathrm{L}$ et que la fenêtre satisfait $2\eta > \mathrm{T}(\mathrm{L}-1)^{-1}$.

Les conditions **A1** et **A2** sont des hypothèses classiques pour la régression non paramétrique. L'hypothèse **A2** est satisfaite par exemple lorsque K est le noyau d'Epanechnikov. La condition **A3** assure que la grille des points de discrétisation est suffisamment fine et que la fenêtre η n'est pas trop petite de façon à ce que l'estimateur soit bien défini.

On écrit $\widetilde{\mu}(t) - \mu(t)$de la manière suivante

$$\widetilde{\mu}(t) - \mu(t) = \frac{\dfrac{1}{\eta \mathrm{L}} \displaystyle\sum_{l=1}^{\mathrm{L}} \mathrm{K}\left(\frac{t-t_l}{\eta}\right)\big(\mu(t_l) - \mu(t)\big)}{\dfrac{1}{\eta \mathrm{L}} \displaystyle\sum_{l=1}^{\mathrm{L}} \mathrm{K}\left(\frac{t-t_l}{\eta}\right)} \tag{73}$$

Comme le noyau K a un support compact, on en déduit que $\mathrm{K}\left(\frac{t_l - t}{\eta}\right) > 0$ seulement si $t_l \in [t-\eta, t+\eta]$. Or par hypothèse les instants d'observation sont équidistants dans $[0,\mathrm{T}]$, *i.e.* $t_l = \mathrm{T}(l-1)/(\mathrm{L}-1),\ l = 1,\ldots,\mathrm{L}$, il y a au plus $2\eta(\mathrm{L}-1)/\mathrm{T}$ valeurs de $\mathrm{K}\left(\frac{t_l - t}{\eta}\right)$ strictement positives et l'hypothèse **A3** empêche tous les termes de (73) d'être nuls. Comme la fonction μ est β-Hölder, on a

$$\begin{aligned}\left|\frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \frac{1}{h} \mathrm{K}\left(\frac{t_l - t}{h}\right)\big(\mu(t_l) - \mu(t)\big)\right| &\leq \frac{1}{\mathrm{L}} \sum_{t_l \in [t-\eta, t+\eta]} \frac{1}{\eta} \mathrm{K}\left(\frac{t_l - t}{\eta}\right) \big|\mu(t_l) - \mu(t)\big| \\ &\leq \Lambda |2\eta|^{\beta} \frac{1}{\mathrm{L}} \sum_{t_l \in [t-\eta, t+\eta]} \frac{1}{\eta} \mathrm{K}\left(\frac{t_l - t}{\eta}\right)\end{aligned}$$

Par l'approximation de la somme de Riemann, comme le noyau K est une fonction continue à support compact, on a, lorsque $\mathrm{L} \to \infty$,

$$\left| \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \frac{1}{\eta} \mathrm{K}\left(\frac{t_l - t}{\eta} \right) - \frac{1}{\eta} \int \mathrm{K}\left(\frac{u-t}{\eta} \right) du \right| \to 0$$

et par le changement de variable $x = (u-t)/\eta$ on obtient $\frac{1}{\eta} \int_{\mathbb{R}} \mathrm{K}\left(\frac{u-t}{\eta} \right) du = \int_{\mathbb{R}} \mathrm{K}(x) dx < +\infty$. On a donc prouvé que la borne donnée dans (5.23) est vraie.

Annexe B : Adaptation de Courbotree et CourboForest au contexte des courbes de consommation électrique et aspects pratiques de l'implémentation

Prédiction de courbes en séparant niveau et forme Lorsque l'on travaille sur des données de consommation électrique, les courbes considérées ont des niveaux extrêmement hétérogènes, et l'algorithme du Courbotree basé sur la distance euclidienne peut mal fonctionner lorsqu'il est appliqué sur les données brutes. Usuellement (parmi les ingénieurs de EDF), on n'utilise donc l'algorithme Courbotree (ou le CourboForest) que sur les *formes* des courbes, obtenues en divisant celles-ci par leur moyenne. Dans notre contexte, on modélise donc séparément la forme des courbes et leur niveau selon un modèle multiplicatif,

$$\mathrm{Y}_i(t) = \check{\mathrm{Y}}_i \mathrm{S}_i(t), \quad i \in \mathrm{U}, \quad t \in [0, \mathrm{T}], \tag{74}$$

avec $\check{\mathrm{Y}}_i = \frac{1}{\mathrm{L}} \sum_{l=1}^{\mathrm{L}} \mathrm{Y}_i(t_l)$ le niveau moyen de la courbe sur les instants de discrétisation et $\mathrm{S}_i = \frac{\mathrm{Y}_i}{\check{\mathrm{Y}}_i}$ sa forme, de moyenne 1 pour toute courbe non toujours nulle. [4]

Puisque nous utilisons ici des techniques de statistique non paramétrique, il pourrait être cohérent d'utiliser une forêt aléatoire pour estimer le niveau moyen $\check{\mathrm{Y}}_i$. Cependant, en pratique, on dispose en général pour chaque unité i d'une variable de consommation totale annuelle C_i déduite de la facturation. Cette variable C_i est en général très corrélée avec le niveau moyen $\check{\mathrm{Y}}_i$ (la seule différence étant la période temporelle considérée) et on suggère donc de modéliser le niveau moyen $\check{\mathrm{Y}}_i$ par un modèle linéaire de la forme

$$\check{\mathrm{Y}}_i = a + b\mathrm{C}_i + e_i, \tag{75}$$

avec e_i un résidu d'espérance nulle. Ce modèle sera estimé par les moindres carrés ordinaires (non pondérés) et, pour chaque unité i, on aura le niveau estimé $\hat{\check{\mathrm{Y}}}_i = \hat{a} + \hat{b}\mathrm{C}_i$.

On estime ensuite un modèle non paramétrique sur les formes des courbes par le Courbotree (arbres de régression pour des courbes) ou le CourboForest, *i.e.*

$$\mathrm{S}_i(t) = \phi(\mathbf{X}_i, t) + \eta_i(t), \tag{76}$$

avec η_i un résidu d'espérance nulle.

4. Par convention, pour une courbe toujours nulle, la forme et le niveau seront nuls.

On en déduit enfin la courbe estimée par une multiplication du niveau et de la forme,

$$\hat{\mathrm{Y}}_i(t) = \hat{\check{\mathrm{Y}}}_i \hat{\mathrm{S}}_i(t).$$

Indicateur d'influence pour les forets aléatoires CourboForest, lorsque l'on dissocie estimation de la forme et du niveau Ici, on cherche à construire des estimateurs robustes de courbes moyennes de domaines par des forêts aléatoires Courboforest dans l'approche basée sur le biais conditionnel. Lorsque, l'on dissocie l'estimation du niveau des courbes $\check{\mathrm{Y}}_i$ de l'estimation de leurs formes S_i, il est nécessaire d'adapter l'indicateur d'influence présenté dans 4.4.1 afin de tenir compte du modèle présenté ci-dessus. L'équation (4.46) de la sous-section devient :

$$\mathrm{I}_{1i}^{\hat{\mu}_d} = \frac{1}{\mathrm{N}_d}\left(\sum_{j\in \mathrm{U}_d - s_d} (\hat{\check{\mathrm{Y}}}_j \hat{\mathrm{S}}_j - \hat{\check{\mathrm{Y}}}_j^{(-i)} \hat{\mathrm{S}}_j^{(-i)})\right), \quad i \in s, \quad d = 1, \ldots \mathrm{D}, \tag{77}$$

avec $\hat{\check{\mathrm{Y}}}_j^{(-i)}$ et $\hat{\mathrm{S}}_j^{(-i)}$ respectivement le niveau et la forme de la courbe de l'unité $j \in \mathrm{U}$ estimés en se basant sur $s^{(-i)}$. Or on a

$$\hat{\check{\mathrm{Y}}}_j \hat{\mathrm{S}}_j - \hat{\check{\mathrm{Y}}}_j^{(-i)} \hat{\mathrm{S}}_j^{(-i)} = (\hat{\check{\mathrm{Y}}}_j - \hat{\check{\mathrm{Y}}}_j^{(-i)})\hat{\mathrm{S}}_j + \hat{\check{\mathrm{Y}}}_j^{(-i)}(\hat{\mathrm{S}}_j - \hat{\mathrm{S}}_j^{(-i)}).$$

On peut donc décomposer cette expression de l'influence en deux parties :

$$\mathrm{I}_{1i}^{\hat{\mu}_d} = \underbrace{\frac{1}{\mathrm{N}_d}\sum_{j\in \mathrm{U}_d - s_d} (\hat{\check{\mathrm{Y}}}_j - \hat{\check{\mathrm{Y}}}_j^{(-i)})\hat{\mathrm{F}}_j}_{\mathrm{I}_{1i}^{\hat{\check{\mathrm{Y}}}_d}} + \underbrace{\frac{1}{\mathrm{N}_d}\sum_{j\in \mathrm{U}_d - s_d} \hat{\check{\mathrm{Y}}}_j^{(-i)}(\hat{\mathrm{S}}_j - \hat{\mathrm{S}}_j^{(-i)})}_{\mathrm{I}_{1i}^{\hat{\mathrm{S}}_d}}, \quad i \in s, \quad d = 1, \ldots \mathrm{D}. \tag{78}$$

Intuitivement, la première partie de cet indicateur représente l'influence de l'unité i sur l'estimation du niveau de la courbe moyenne du domaine d alors que la seconde représente son influence sur l'estimation de son niveau. On va donc proposer un indicateur pour chacun de ces deux termes et ensuite les sommer pour obtenir l'indicateur global.

Le niveau des courbes non échantillonnées $\hat{\check{\mathrm{Y}}}_j$, $j \in \mathrm{U} - s$ est estimé par une régression linéaire simple (voir équation 75) , donc l'influence de $i \in s$ sur l'estimation du niveau moyen de la population $\frac{\hat{\check{\mathrm{Y}}}_d}{\mathrm{N}_d}$ $d = 1, \ldots \mathrm{D}$ est donné par (voir sous-section 4.4.1)

$$\frac{1}{\mathrm{N}_d}\sum_{j\in \mathrm{U}_d - s_d} (\hat{\check{\mathrm{Y}}}_j - \hat{\check{\mathrm{Y}}}_j^{(-i)})\hat{\mathrm{S}}_j = \frac{1}{\mathrm{N}_d}\left(\sum_{j\in \mathrm{U}_d - s_d} \mathrm{C}_j \hat{\mathrm{S}}_j\right)(\hat{b} - \hat{b}^{(-i)}) \tag{79}$$

$$\mathrm{I}_{1i}^{\hat{\check{\mathrm{Y}}}_d} = \frac{1}{\mathrm{N}_d}\left(\sum_{j\in \mathrm{U}_d - s_d} \mathrm{C}_j \hat{\mathrm{S}}_j\right)\frac{\mathrm{C}_i(\mathrm{Y}_i - b\mathrm{C}_i)}{\sum_{j\in s}\mathrm{C}_j^2}. \tag{80}$$

Pour obtenir l'indicateur souhaité, on remplace finalement la quantité inconnue b par son estimateur $\hat{b}$:

$$\mathrm{I}_{1i}^{\hat{\check{\mathrm{Y}}}_d} = \frac{1}{\mathrm{N}_d}\left(\sum_{j\in \mathrm{U}_d - s_d} \mathrm{C}_j \hat{\mathrm{S}}_j\right)\frac{\mathrm{C}_i(\mathrm{Y}_i - \hat{b}\mathrm{C}_i)}{\sum_{j\in s}\mathrm{C}_j^2}. \tag{81}$$

Pour construire le second terme $\mathrm{I}_{1i}^{\hat{\mathrm{S}}_d}, \quad i \in s, \quad d = 1, \ldots \mathrm{D}$, on procède d'une manière similaire à celle employée dans la Sous-Section 4.4.1 en régressant les prédictions obtenues pour chaque arbre sur le nombre de présences de chaque unité dans l'échantillon sur lequel cet arbre a été construit. Toutefois, on modifie légèrement la démarche précédente pour coller à l'expression de l'équation (78). En effet, notons $\hat{\mu}_d^{a;\mathrm{S}}$ la forme moyenne des unités non échantillonnées pondérées par $\hat{\hat{\mathrm{Y}}}_j^{(-i)}$ (leur niveau estimé sans l'unité i) :

$$\hat{\mu}_d^{a;\mathrm{S}} = \frac{1}{\mathrm{N}_d} \sum_{j \in \mathrm{U}_d - s_d} \hat{\mathrm{F}}_j \hat{\hat{\mathrm{Y}}}_j^{(-i)}, \quad i \in s, \quad d = 1, \ldots \mathrm{D}. \tag{82}$$

Pour chaque domaine d et chaque unité i, on régresse cette quantité sur le nombre de présences de l'unité i dans l'arbre a :

$$\hat{\mu}_d^{a;\mathrm{S}} = \mathrm{A}_{di}^{\mathrm{S}} + \mathrm{B}_{di}^{\mathrm{S}} n_i^a + \mathrm{E}_{di}^{a;\mathrm{S}}, \quad i \in s, \quad d = 1, \ldots \mathrm{D},$$

avec $\mathrm{E}_{di}^{a;\mathrm{S}}$ un résidu d'espérance nulle. L'indicateur d'influence recherché est finalement l'estimateur des moindres carrés ordinaires du cœfficient de cette régression : $\mathrm{I}_{1i}^{\hat{\mathrm{S}}_d} = \hat{\mathrm{B}}_{di}^{\mathrm{S}}$ Comme précédemment, pour les unités i dont le nombre de sélections n_i^a est constant pour l'ensemble des arbres a, on prend $\mathrm{I}_{1i}^{\hat{\mathrm{S}}_d} = 0$.

On a finalement l'indicateur d'influence global

$$\mathrm{I}_{1i}^{\hat{\mu}_d} = \mathrm{I}_{1i}^{\hat{\mathrm{S}}_d} + \mathrm{I}_{1i}^{\hat{\hat{\mathrm{Y}}}_d}.$$

On peut ensuite l'utiliser pour construire des estimateurs robustes par l'équation (4.31).

Bibliographie

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106(494) :534–543. 36

Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J.-M. (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*, 11(01) :1350003. 18

Ardilly, P. (2006). *Les techniques de sondage*. Editions Technip. 13, 26

Ardilly, P. (2014). Estimation régionale de taux de pauvreté utilisant une technique de calage. In *Actes du 8e colloque francophone sur les sondages, Dijon*. 101, 108, 109, 136

Ardilly, P. and Tillé, Y. (2006). *Sampling methods : Exercises and solutions*. Springer Science & Business Media. 151

Bar-Hen, A. and Poggi, J.-M. (2016). Influence measures and stability for graphical models. *Journal of Multivariate Analysis*, 147 :145–154. 11, 121

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401) :28–36. 10, 100, 101, 105, 106

Beaumont, J.-F. and Bissonnette, J. (2011). Variance estimation under composite imputation : The methodology behind sevani. *Survey Methodology*, 37(171-179). 11, 140, 157, 158, 159, 163, 180

Beaumont, J.-F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian J. of Statistics*, 37(400-416). 140, 159, 160, 162

Beaumont, J.-F., Haziza, D., and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100(3) :555–569. 3, 5, 8, 9, 39, 40, 43, 44, 45, 48, 51, 56, 62, 95, 117, 118, 120

Beaumont, J.-F. and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to poisson sampling. *International Statistical Review*, 80(1) :127–148. 36

Bertail, P. and Combris, P. (1997). Bootstrap généralisé d'un sondage. *Annales d'économie et de statistique*, pages 49–83. 35, 36, 66, 67

Booth, J. G., Butler, R. W., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89(428) :1282–1289. 35, 36, 66

Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5–32. 113

Breiman, L. et al. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, 26(3) :801–849. 113

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press. 111

Brent, R. P. (2013). *Algorithms for minimization without derivatives*. Courier Corporation. 58

Brown, B. (1983). Statistical uses of the spatial median. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 25–30. 49, 50

Burck, L. and Pfeffermann, D. (1990). Robust small area estimation combining time series and cross-sectional data. *Surv. Methodol*, 16 :217–237. 100

Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140(1) :75–91. 8, 18, 25, 51, 106

Cardot, H., De Moliner, A., and Goga, C. (2015). Estimating with kernel smoothers the mean of functional data in a finite population setting. a note on variance estimation in presence of partially observed trajectories. *Statistics & Probability Letters*, 99 :156–166. 140, 141

Cardot, H., Degras, D., and Josserand, E. (2013a). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, 19(5A) :2067–2097. 15, 25, 30, 34, 147, 159

Cardot, H., Dessertaine, A., Goga, C., Josserand, É., and Lardin, P. (2013b). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data : An illustration on electricity consumption. *Survey Methodology*, 39(2) :283–301. 8, 21, 29, 30, 34, 96, 103

Cardot, H., Goga, C., and Lardin, P. (2013c). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7 :562–596. 25, 26, 30, 31, 34

Cardot, H. and Josserand, E. (2011). Horvitz-Thompson estimators for functional data : asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98(1) :107–118. 8, 15, 25, 28

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93(2) :255–268. 102

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396) :1063–1069. 37, 39, 124

Chaouch, M. and Goga, C. (2012). Using complex surveys to estimate the l1-median of a functional variable : Application to electricity load curves. *International Statistical Review*, 80(1) :40–59. 51, 64

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434) :862–872. 50

Chauvet, G. (2007). *Méthodes de bootstrap en population finie*. PhD thesis, ENSAI. 35, 36

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16 :113–131. 140, 154

Chen, T. and Guestrin, C. (2016). Xgboost : A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM. 137

Cochran, W. G. (1977). *Sampling Techniques : 3d Ed.* Wiley. 28

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge University Press Cambridge. 113

Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components : the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1) :206–226. 49

Cuesta-Albertos, J. A., Fraiman, R., and Ransford, T. (2006). Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society*, 37(4) :477–501. 54

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147 :1–23. 9, 13, 140

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function : some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1) :136–154. 14, 18

De'Ath, G. (2002). Multivariate regression trees : a new technique for modeling species–environment relationships. *Ecology*, 83(4) :1105–1117. 112

Degras, D. (2014). Rotation sampling for functional data. *Statistica Sinica*, pages 1075–1095. 23

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. In *Annales de l'INSEE*, pages 3–101. JSTOR. 14, 18

Deville, J. C. (1999). Variance estimation for complex statistics and estimators : Linearization and residual techniques. *Survey methodology*, 25(2) :193–204. 34, 149

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418) :376–382. 30, 109

Dorfman, A. (1994). Open questions in the application of smoothing methods to finite population inference. *COMPUTING SCIENCE AND STATISTICS*, pages 201–201. 33

Efron, B. (1979). Bootstrap methods : another look at the jackknife. *The annals of Statistics*, pages 1–26. 35

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3) :254–261. 21, 108

Favre-Martinoz, C. (2015). *Estimation robuste en population finie et infinie*. PhD thesis, Université Rennes 1. 37

Fay III, R. E. and Herriot, R. A. (1979). Estimates of income for small places : an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a) :269–277. 100

Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics*, 28(1) :51–60. 125

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis : theory and practice*. Springer. 9, 13, 14, 20, 140

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2) :419–440. 54

Gervini, D. (2008). Robust functional estimation using the spatial median and spherical principal components. *Biometrika*, 95 :587–600. 49, 50

Gervini, D. (2012). Outlier detection and trimmed estimation for general functional data. *Statistica Sinica*, 22 :1639–1660. 54

Godambe, V. and Sprott, D. (1971). *Foundations of Statistical Inference : Proceedings of the Symposium on the Foundations of Statistical Inference Prepared Under the Auspices of the René Descartes Foundation and Held at the Department of Statistics, University of Waterloo, Ont., Canada, from March 31 to April 9, 1970*. Holt McDougal. 31

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, volume 1814184. American Statistical Association Ithaca, NY. 35

Gwet, J.-P. and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87(420) :1174–1182. 39, 62

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York. 154

Hajjem, A., Bellavance, F., and Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6) :1313–1328. 137

Hall, P., Müller, H.-G., and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, pages 1493–1517. 18

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics : the approach based on influence functions*, volume 114. John Wiley & Sons. 40, 41, 42

Hart, J. (2013). *Nonparametric smoothing and lack-of-fit tests*. Springer Science & Business Media. 143

Hart, J. D. and Wehrly, T. E. (1993). Consistency of cross-validation when the data are curves. *Stochastic processes and their applications*, 45(2) :351–361. 147

Haziza, D. (2005). Inférence en présence d'imputation simple dans les enquêtes : un survol. *Journal de la société française de statistique*, 146(4) :69–118. 140

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Sample surveys : design, methods and applications*, volume 29 of *Handbook of Statist.*, pages 215–246. Elsevier/North-Holland, Amsterdam. 8, 139, 154

Holmberg, A. (1998). A bootstrap approach to probability proportional-to-size sampling. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 378–383. 36

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*. Springer Series in Statistics. Springer, New York. 13, 14, 140

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260) :663–685. 103

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1) :73–101. 40, 41, 42, 43

Hyndman, R. J. and Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1). 56

Hyndman, R. J. and Ullah, S. (2007). Robust forecasting of mortality and fertility rates : a functional data approach. *Computational Statistics & Data Analysis*, 51(10) :4942–4956. 54

Jiongo, V. D., Haziza, D., and Duchesne, P. (2013). Controlling the bias of robust small-area estimators. *Biometrika*, 100(4) :843–858. 11, 102, 119, 120

Kemperman, J. (1987). The median of a finite measure on a banach space. *Statistical data analysis based on the L1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. 50

Kokic, P. and Bell, P. (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, 10 :419–419. 39

Lardin, P. (2012). *Estimation de synchrones de consommation électrique par sondage et prise en compte d'information auxiliaire*. PhD thesis, Université de Bourgogne. 3, 5, 30, 31

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553) :436–444. 137

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999). Robust principal component analysis for functional data. *Test*, 8 :1–73. 10, 20, 49

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486) :718–734. 54

Lundström, S. and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of official statistics*, 15(2) :305. 144

Mallat, S. (1999). *A wavelet tour of signal processing*. Academic press. 10, 18

Moreno-Rebollo, J., Muñoz-Reyes, A., and Muñoz-Pichardo, J. (1999). Miscellanea. influence diagnostic in survey sampling : conditional bias. *Biometrika*, 86(4) :923–928. 8, 39

Muñoz-Pichardo, J., Munoz-Garcia, J., Moreno-Rebollo, J., and Pino-Mejias, R. (1995). A new approach to influence analysis in linear models. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 393–409. 8, 39, 42

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*. 24

Neyman, J. (1934). On the two different aspects of the representative method : the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4) :558–625. 28

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(1) :265–286. 100

Pfeffermann, D. and Rao, C. R. (2009). *Handbook of Statistics_29A : Sample Surveys : Design, Methods and Applications*, volume 29. Elsevier. 39

Ramsay, J.-O. and Silverman, B.-W. (2005). *Functional Data Analysis*. Springer Series in Statistics, New York, second edition. 9, 13, 14, 15, 19, 20, 21, 140

Rao, J. N., Sinha, S. K., and Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *Canadian Journal of Statistics*, 42(1) :126–141. 102

Rao, J. N. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the american statistical association*, 83(401) :231–241. 35

Rao, J. N. K. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons. 8, 100, 106, 107, 110, 119, 124

Rao, J. N. K. and Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, 22(4) :511–528. 100

Richardson, A. M. and Welsh, A. H. (1995). Robust restricted maximum likelihood in mixed linear models. *Biometrics*, pages 1429–1439. 125

Rivest, L.-P. (1994). Statistical properties of winsorized means for skewed distributions. *Biometrika*, 81(2) :373–383. 39

Royall, R. M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71(355) :657–664. 33

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, pages 581–592. 142

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18(2) :241–252. 31, 101, 109, 110, 148

Särndal, C.-E. and Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley Series in Survey Methodology. John Wiley & Sons, Ltd., Chichester. 8, 139

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer. 8, 13, 29, 34

Segal, M. and Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(1) :80–87. 114

Shao, J. (2009). Nonparametric variance estimation for nearest neighbor imputation. *Journal of Official Statistics*, 25 :55–62. 162

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445) :254–265. 148

Shao, J. and Tu, D. (2012). *The jackknife and bootstrap*. Springer Science & Business Media. 35

Sinha, S. K. and Rao, J. (2009). Robust small area estimation. *Canadian Journal of Statistics*, 37(3) :381–399. 11, 102, 115, 123, 126

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87(419) :755–765. 35

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277. 50

Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.*, 93(444) :1403–1418. 140, 142, 143

Stéphan, V. and Cogordan, F. (2009). Courbotree : Application des arbres de regression multivaries pour la classification de courbes. *Revue MODULAD juin*. 11, 101, 112, 136

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2). 56

Tillé, Y. (2001). Théorie des sondages. *Dunod, Paris*. 13, 23, 26

Toth, D. and Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496) :1626–1636. 111

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531. 54

Tzavidis, N., Marchetti, S., and Chambers, R. (2010). Robust estimation of small-area means and quantiles. *Australian & New Zealand Journal of Statistics*, 52(2) :167–186. 102

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite population sampling and inference : a prediction approach*. John Wiley. 11, 32, 33, 101, 105, 110, 111

Vardi, Y. and Zhang, C.-H. (2000). The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4) :1423–1426. 50

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London. 143

Weiszfeld, E. (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43 :355–386. 50

Welsh, A. H. and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 60(2) :413–428. 39

Ye, J., Chow, J.-H., Chen, J., and Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 2061–2064. ACM. 137